Beyond Uniform Detection: Adaptive Hallucination Detection for RAG Across Response Regimes

Jungwuk Park[1*†] **Sejong Ryu**[1*] **Jy-yong Sohn**[2] **Jaekyun Moon**[1]
[1]KAIST [2]Yonsei University
savertm9@gmail.com, ryusejong@kaist.ac.kr, jysohn1108@yonsei.ac.kr, jmoon@kaist.edu

Abstract

Existing hallucination detectors for retrieval-augmented generation (RAG), whether based on model outputs (e.g., likelihood) or internal representations, typically apply a uniform detection strategy across responses. However, we find that different detection signals are informative in different response regimes, making a uniform strategy insufficient for both short- and long-form generation. In short-form tasks such as factoid QA, hallucination typically appears as selecting an incorrect final answer among multiple context-relevant candidates. In this regime, model-output-based signals that discriminate among plausible answers are particularly important. We therefore propose a bidirectional likelihood measure that evaluates logical consistency based on the retrieved evidence and a reasoning step generated after the response. In contrast, long-form descriptive responses are more prone to gradual drift: as generation proceeds, the model increasingly relies on its own prior generations and parametric knowledge, while unsupported continuations may remain locally plausible, making output-based cues less informative. For this regime, we measure context–knowledge conflict by tracking the directional alignment of context-grounding versus internal-knowledge contributions in the hidden states. Building on these regime-specific signals, we introduce ARGUS, an adaptive hallucination detector that emphasizes the more effective signal according to the response regime. Experiments on multiple RAG benchmarks show that our method achieves strong performance across both short- and long-form settings.

1 Introduction

Large language models (LLMs) are prone to hallucination, often generating factually incorrect or unsupported responses that nevertheless appear fluent and plausible [11, 10]. Retrieval-augmented generation (RAG) mitigates this issue by grounding responses in retrieved documents [16, 24, 6], yet hallucinations still arise even when relevant evidence is provided [20, 9, 26]. Detecting such failures is therefore essential for the reliable deployment of RAG systems, especially in high-stakes settings.

Recent work has shown that when LLMs are provided with external context, hallucinations often arise from conflicts between the provided context and the model's parametric knowledge [29, 17], and recent detectors for RAG have followed this perspective by measuring such conflict in internal representations [26, 31]. Alternative output-based cues have also been widely studied in the broader hallucination detection literature, including likelihoods [23, 18, 33], cross-sample consistency [19, 2], and verbalized self-checks [14, 9]. However, we find that different detection signals are informative in different response regimes, suggesting that a uniform detection strategy is insufficient.

In short-form tasks such as factoid QA, hallucination often appears as selecting an incorrect answer among multiple context-relevant candidates, while the response remains broadly grounded in the

*Equal contribution. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

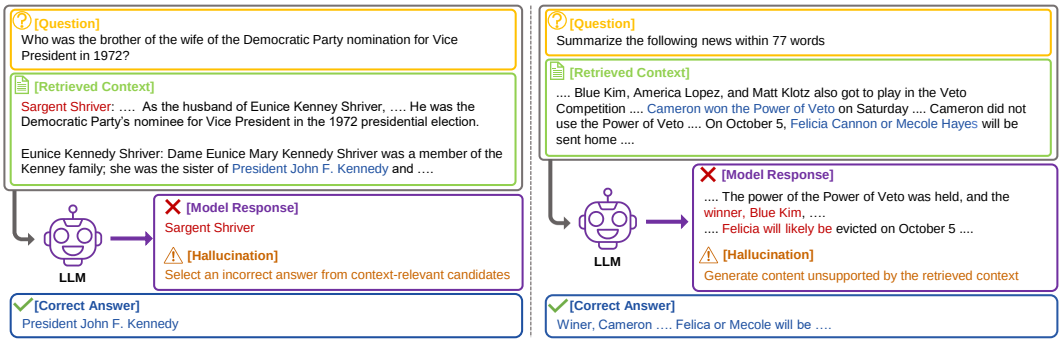


Figure 1: Hallucination in RAG varies across response regimes. (Left) In short-form QA, the model selects an incorrect answer among candidates in the retrieved context. (Right) In long-form generation, the response starts grounded in the retrieved evidence but gradually drifts into unsupported content.

retrieved evidence (Figure 1, left). In this regime, conflict between retrieved context and internal knowledge can be subtle, since the model may remain anchored in the evidence while making an incorrect choice; as a result, internal-representation-based detectors for RAG, such as ReDeEP [26] and LUMINA [31], which rely on detecting context–knowledge conflict, struggle here. In contrast, simple likelihood-based methods (e.g., Perplexity [23]) remain relatively effective (Figure 2).

In contrast, long-form descriptive responses are more vulnerable to gradual drift, where the response departs from the retrieved evidence during generation, as shown in Figure 1 (right). As generation proceeds, the model increasingly conditions on its own previously generated tokens, while parametric knowledge may play a larger role in sustaining fluent continuation beyond the retrieved evidence. As a result, unsupported continuations may remain locally plausible, making likelihood-based cues such as Perplexity less informative in this regime. Conversely, internal-representation-based detectors such as ReDeEP and LUMINA, which capture context–knowledge conflict, are more effective here (Figure 2). This signal reversal motivates our regime-dependent view: no single detector class consistently dominates, and different signals become informative in different settings.

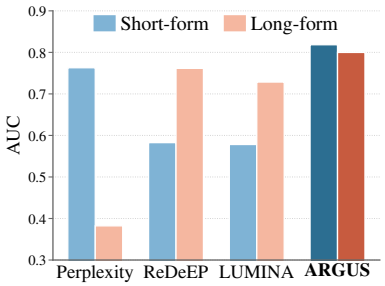


Figure 2: Detection performance in short- and long-form RAG settings.

Motivated by this, we propose two new hallucination signals tailored to each regime. For short-form responses, we introduce a *bidirectional likelihood* score that combines the forward generation likelihood with a backward consistency check based on a reasoning step generated after the response. For long-form responses, we propose a *hidden-state conflict* measure that tracks the directional alignment of context-grounding versus internal-knowledge contributions in the hidden-state space. Building on these signals, we introduce ARGUS, an *adaptive* hallucination detector that uses response length as a simple yet effective proxy to emphasize the more appropriate signal, rather than applying a uniform detection strategy. Across 8 diverse datasets spanning short- and long-form RAG settings, our adaptive detector achieves strong overall performance across both regimes (Figure 2).

Contributions. **(1) Regime-dependent signal reversal.** To our knowledge, we are the first to identify and empirically characterize a regime-dependent reversal in RAG hallucination detection: likelihood-based signals are more informative for short-form, whereas hidden-state conflict signals are more informative for long-form. **(2) Regime-specialized detection.** Motivated by this reversal, we design two regime-tailored signals: bidirectional likelihood scoring for short-form detection and hidden-state conflict measurement via attention-based directional alignment for long-form detection, and combine them into an adaptive detector via length-based weighting. **(3) Strong empirical results.** We demonstrate that the proposed method achieves strong overall performance across 8 diverse datasets in short- and long-form settings.

2 Related Work

Hallucination detection from model outputs. One line of work detects hallucination at the output level, without accessing internal representations. *Uncertainty-based* methods estimate hallucination

risk from per-token probabilities: perplexity [23] averages negative log-likelihoods across tokens, while length-normalized entropy [18] averages per-token entropies. Focus [33] refines this by restricting the aggregation to content-bearing tokens (e.g., named entities) that are more diagnostic of hallucination. *Consistency-based* methods such as SelfCheckGPT [19] sample multiple responses and measure their consistency, but require multiple generations, introducing computational overhead. *Verbalization-based* methods prompt an LLM to judge its own faithfulness: P(True) [14] queries the model's self-assessed confidence, and RefChecker [9] performs reference-based claim verification.

Hallucination detection from internal representations. A complementary line of work looks inside the model to detect hallucination from internal representations. The logit lens [21] projects intermediate hidden states into vocabulary space for layer-wise prediction inspection. [32] shows that attention patterns encode constraint-satisfaction signals useful for detecting factual errors; [1] trains probes on hidden states to detect false statements; and another approach uses eigenvalue analysis of internal representations to identify hallucinations either across multiple samples [2] or from a single forward pass without any learned classifier [25].

In the RAG setting, RAGTruth [20] examines hallucinations that arise in context-grounded settings and contributes the first high-quality, manually annotated dataset for their systematic study. Recent studies attribute these hallucinations to conflicts between the provided evidence and the model's parametric knowledge [29, 17], motivating detection methods that separately measure the model's utilization of external context and its reliance on parametric knowledge. Specifically, ReDeEP [26] measures these two signals via attention-based context scores and FFN-based knowledge scores (computed via logit lens), though it relies on selecting specific attention heads and transformer layers, which may require model-specific tuning [31]. LUMINA [31] addresses this by measuring both signals in a layer-agnostic manner using distributional distance against a random document and a logit-lens-based information processing rate, yielding improved results.

Despite these advances, existing approaches apply a single detection mechanism uniformly, regardless of the response regime. Our work addresses this gap with two regime-specialized signals adaptively applied to each regime. For short-form responses, we introduce a bidirectional likelihood that evaluates the response against a reasoning step generated after it, capturing inconsistencies that prior likelihood-based methods miss. For long-form responses, we introduce a hidden-state conflict measure that contrasts context and internal-knowledge contributions via attention decomposition in the hidden-state space, rather than the logit-lens-based utilization signals used by prior approaches.

3 Methods

In this work, we focus on detecting hallucinations in RAG settings where the retrieved context is accurate and relevant, as in [26, 31]. These hallucinations can manifest differently across response regimes: in short-form responses, the model may select an incorrect answer among multiple context-relevant candidates; in long-form responses, generation may gradually drift from the retrieved evidence. To address this, we propose ARGUS, an adaptive detection framework with two regime-specialized signals. First, we introduce a bidirectional likelihood score that evaluates the response likelihood given the preceding evidence and a reasoning step generated after it, enabling effective discrimination among plausible candidates in short-form tasks (Section 3.1). Second, we propose a hidden-state conflict score that contrasts context-grounding and internal-knowledge contributions, designed to capture gradual drift in long-form generation (Section 3.2). The two signals are adaptively weighted by response length, emphasizing the more effective signal in each regime (Section 3.3). An overview of ARGUS is shown in Figure 3 and summarized in Algorithm 1 of Appendix L.

3.1 Likelihood Score with Bidirectional Consistency for Short-Form Detection

In short-form RAG tasks such as factoid or extractive QA, the retrieved context often contains multiple related candidates, and hallucination typically arises when the model selects the wrong one with high confidence. The key challenge is to distinguish the correct answer from plausible alternatives. We further support this characterization empirically: across short-form QA datasets, an average of $\sim$86% of hallucinated responses appear in the context under normalized lexical matching, suggesting candidate misselection rather than fully ungrounded generation (Appendix D). We therefore propose a likelihood-based score that goes beyond simple response likelihood by generating a reasoning sequence after the response and checking the consistency between evidence, response, and reasoning.

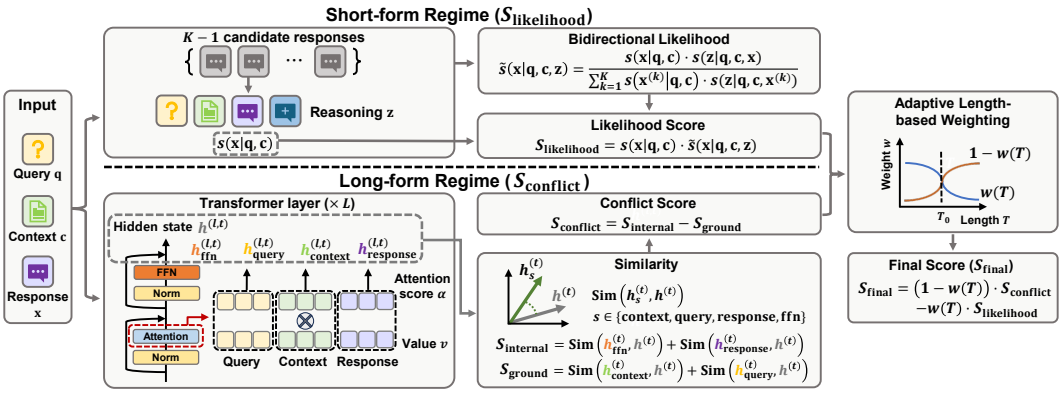


Figure 3: Overview of our proposed ARGUS method. In short-form tasks, we compute a bidirectional likelihood that evaluates the response against a reasoning step generated after it. In long-form tasks, we compute a hidden-state conflict score from per-layer context-grounding vs. internal-knowledge contributions. The two are adaptively combined via length-based weighting $w(T)$.

Forward likelihood. A natural starting point is the model's own confidence in its response: if the model assigns low probability to the generated tokens, this already signals uncertainty. Let $\mathbf{q}$ denote the query, $\mathbf{c}$ the retrieved context, and $\mathbf{x} = (x_1, \ldots, x_T)$ the entire generated response. We measure this confidence via the geometric mean of token-level probabilities:

$$s(\mathbf{x} \mid \mathbf{q}, \mathbf{c}) = \left(\prod_{i=1}^{T} p(x_i \mid \mathbf{q}, \mathbf{c}, x_{<i}) \right)^{\frac{1}{T}}, \tag{1}$$

where s is a length-normalized likelihood score. However, confidence alone is insufficient: the model can select a wrong but plausible candidate with high confidence, a *confidently wrong* failure mode.

Bidirectional likelihood. To address this, we introduce a backward consistency check based on reasoning. The intuition is that a correct response is more likely to elicit reasoning consistent with the evidence, while a hallucinated one tends to produce less consistent reasoning due to weak grounding in the evidence, providing an additional signal to discriminate among context-relevant candidates. Appendix E.1 examines this intuition empirically, showing that the generated reasoning favors an alternative candidate more often for hallucinated than for faithful responses.

Given the original response $\mathbf{x}$, we generate a short reasoning sequence $\mathbf{z}$ within a limited token budget by greedy decoding conditioned on $(\mathbf{q}, \mathbf{c}, \mathbf{x})$, prompting the model to reason over its response. Although $\mathbf{z}$ is generated after $\mathbf{x}$, we condition $\mathbf{x}$ on $\mathbf{z}$ to re-evaluate its likelihood, a backward-looking check that complements the original forward generation, hence the term *bidirectional*. We perform this re-evaluation by comparing the response against plausible alternatives.

Concretely, we construct K candidate responses $\{\mathbf{x}^{(k)}\}_{k=1}^{K}$ from the top-K candidates for the first response token x_1, completing each via greedy decoding (with $\mathbf{x}^{(1)} = \mathbf{x}$ denoting the original response). Since short-form responses are typically only a few tokens and K is small ($K = 3$ in our experiments), the additional decoding cost remains manageable. We multiply each candidate's forward confidence $s(\mathbf{x}^{(k)} \mid \mathbf{q}, \mathbf{c})$ and reasoning likelihood $s(\mathbf{z} \mid \mathbf{q}, \mathbf{c}, \mathbf{x}^{(k)})$, reusing the same $\mathbf{z}$ generated from $\mathbf{x}$. This product favors candidates that are evidence-plausible and reasoning-compatible. Normalizing across the candidates, the bidirectional likelihood score is defined as

$$\tilde{s}(\mathbf{x} \mid \mathbf{q}, \mathbf{c}, \mathbf{z}) = \frac{s(\mathbf{x} \mid \mathbf{q}, \mathbf{c}) \cdot s(\mathbf{z} \mid \mathbf{q}, \mathbf{c}, \mathbf{x})}{\sum_{k=1}^{K} s(\mathbf{x}^{(k)} \mid \mathbf{q}, \mathbf{c}) \cdot s(\mathbf{z} \mid \mathbf{q}, \mathbf{c}, \mathbf{x}^{(k)})}. \tag{2}$$

Intuitively, $\tilde{s}$ assigns a higher score when the original response stands out among the top-K candidates under the reasoning consistency check.

Combined score. The final output-level score combines both signals:

$$\mathcal{S}_{\text{likelihood}} = s(\mathbf{x} \mid \mathbf{q}, \mathbf{c}) \cdot \tilde{s}(\mathbf{x} \mid \mathbf{q}, \mathbf{c}, \mathbf{z}). \tag{3}$$

The first term, $s(\mathbf{x} \mid \mathbf{q}, \mathbf{c})$, captures the model's forward confidence in the response, while the second term, $\tilde{s}$, is a normalized comparative score over the top-K candidates incorporating reasoning

consistency. This second term is useful for distinguishing among context-relevant candidates, as the reasoning step provides a stronger consistency check. However, the reasoning-conditioned term may align with the proposed answer even when it is incorrect; retaining the forward likelihood as a direct evidence-grounded signal mitigates this risk. The product encourages a high score when both the forward confidence and the reasoning consistency signal favor the response, and drops when either signal weakens (see Appendix E for the complementary behavior of the two terms).

3.2 Hidden-State Conflict via Source Decomposition for Long-Form Detection

In long-form generation, hallucination typically emerges as a gradual drift, with the response progressively deviating from the retrieved evidence. The drift is hard to detect at the output level: each generated token remains locally fluent and probable, so token-level likelihood stays high even as the response globally drifts. In this regime, the conflict between retrieved context and the model's internal knowledge becomes a more informative signal. We therefore examine this conflict at the representation level, where the hidden states reflect the relative contributions of evidence and internal knowledge, making it easier to characterize even when the surface output remains plausible.

In a transformer, the hidden state at each generation step is formed through cumulative residual updates from attention and FFN layers. The attention output at each layer is a weighted sum over attended positions, which allows us to separately analyze the contributions of different source positions (e.g., query or context). The FFN output, in turn, has been associated with encoding parametric knowledge by storing learned patterns in the FFN's weights in prior interpretability work [7, 3], motivating its use as part of the model's internal contribution.

Key idea. We partition attention positions into three regions (the retrieved context, the user query, and the model's own previously generated tokens), yielding four components, including the FFN output. We group these into externally provided conditioning signals (the retrieved context and the user query, both supplied to the model as input) and internal contributions (prior response tokens and the FFN, reflecting autoregressive self-conditioning and parametric knowledge). We assess each component's directional alignment with the final hidden state via cosine similarity. We then compare internal and external alignment as our drift signal: stronger internal alignment suggests that generation is becoming more internally driven and thus more likely to drift from the retrieved evidence.

Attention decomposition. In a pre-norm transformer, where layer normalization is applied at the input of each attention or FFN block, attention and FFN outputs are added to the residual stream at each layer. We track source-specific accumulated updates and compare them with the final hidden state in the same normalized space.

Let $h^{(t)} \in \mathbb{R}^d$ denote the final hidden state of the t-th generated token used for next-token prediction (the output after the model's final-layer normalization), and let $r^{(t)}$ denote the residual stream just before this normalization. Let $\mathcal{Q}$, $\mathcal{C}$, and $\mathcal{R}_{<t}$ denote the position sets for the query, retrieved context, and prior response tokens. In layer l, restricting the attention output to positions in $\mathcal{P}_{\mathrm{s}}$ gives the per-source per-layer attention contribution:

$$h_{\mathrm{s}}^{(l,t)} = \left(\sum_{i \in \mathcal{P}_{\mathrm{s}}} \alpha_i^{(l,t)} v_i^{(l,t)} \right) W_O^{(l)}, \qquad \mathrm{s} \in \{\mathrm{context}, \mathrm{query}, \mathrm{response}\},$$

where $\alpha_i^{(l,t)}$, $v_i^{(l,t)}$, and $W_O^{(l)}$ are the attention weight at position i, the value vector, and the output projection (multi-head index omitted for clarity); and the per-layer FFN output is $h_{\mathrm{ffn}}^{(l,t)}$.

Aggregating each source's per-layer updates across layers and applying the model's final-layer normalization to the result, with the same global statistics as those of $r^{(t)}$, gives

$$h_{\mathrm{s}}^{(t)} = \mathrm{Norm}\left(\sum_{l=1}^{L} h_{\mathrm{s}}^{(l,t)} \right), \qquad \mathrm{s} \in \{\mathrm{context}, \mathrm{query}, \mathrm{response}, \mathrm{ffn}\},$$

where Norm denotes the model's final-layer normalization, the same one that produces $h^{(t)}$ from $r^{(t)}$. This source-wise construction is motivated by the additive structure of the pre-norm residual stream; details and discussion are in Appendix G. For each source, the cosine similarity between $h_{\mathrm{s}}^{(t)}$ and $h^{(t)}$ serves as an interpretive indicator of how strongly that source aligns with the realized representation.

Conflict score. Based on these components, we define two scores for each token t using cosine similarity $\mathrm{Sim}(a,b) = \frac{a^\top b}{\|a\|\,\|b\|}$. The context-grounding score is

$$S^{(t)}_{\text{ground}} = \mathrm{Sim}(h^{(t)}_{\text{context}}, h^{(t)}) + \mathrm{Sim}(h^{(t)}_{\text{query}}, h^{(t)}), \tag{4}$$

which measures how strongly the final hidden state aligns with the external inputs (the retrieved context and the user query), capturing the degree of context grounding. The internal-knowledge score is

$$S^{(t)}_{\text{internal}} = \mathrm{Sim}(h^{(t)}_{\text{ffn}}, h^{(t)}) + \mathrm{Sim}(h^{(t)}_{\text{response}}, h^{(t)}), \tag{5}$$

which measures alignment with the internal signals.

We define the hidden-state conflict score for each token as $\mathcal{S}^{(t)}_{\text{conflict}} = S^{(t)}_{\text{internal}} - S^{(t)}_{\text{ground}}$, and average over all response tokens:

$$\mathcal{S}_{\text{conflict}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{S}^{(t)}_{\text{conflict}}. \tag{6}$$

When generation remains faithful to the evidence, $S^{(t)}_{\text{ground}}$ tends to dominate, keeping the conflict score relatively low; as the model increasingly relies on its own prior outputs or parametric knowledge, $S^{(t)}_{\text{internal}}$ grows, pushing the score upward.

We further support this drift hypothesis empirically (Figure 4; detailed analysis across backbones in Appendix F.1) by partitioning each long-form response into three positional segments (front, middle, back) and plotting the mean $\mathcal{S}_{\text{conflict}}$ in each segment. The hallucinated curve lies consistently above the faithful one and increases more toward later segments, indicating a stronger shift away from context grounding as the response grows longer, supporting $\mathcal{S}_{\text{conflict}}$.

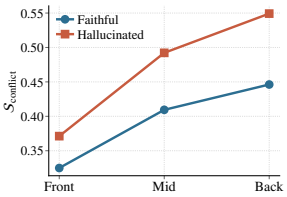


Figure 4: Mean conflict score across response segments.

Comparison with ReDeEP [26] and LUMINA [31]. ReDeEP and LUMINA also aim to detect RAG hallucinations by contrasting retrieved-context and internal-knowledge signals. The main difference lies in how this contrast is quantified. Prior methods estimate these signals separately, using attention/FFN components or predictive distributions, and then infer hallucination from their combination; ReDeEP additionally relies on model-specific component selection, while LUMINA requires an extra forward computation with a random document for distributional comparison. In contrast, our method captures this conflict more directly in hidden-state space by decomposing token-level attention into source-specific contributions, combining them with the FFN contribution, and comparing their alignment to the final hidden state. This strategy yields strong overall empirical performance on long-form benchmarks (Section 4.2).

3.3 Adaptive Length-Based Weighting

Although both signals are defined for any response length, they target different patterns across response regimes. In short responses, the model produces a brief answer in a few tokens; the key is whether that answer is supported by both the evidence and the model's reasoning, which the bidirectional likelihood is designed to assess. In long responses, each token may receive high likelihood, yet the response as a whole may gradually drift from the evidence; this drift is difficult to capture with likelihood-based scores but can be detected by the hidden-state conflict signal, which tracks the shifting balance between context grounding and internal knowledge as generation proceeds.

To combine these signals adaptively, we use response length as a practical proxy. Length is not intended to define the response regime by itself; rather, it serves as an observable proxy for the extent of autoregressive self-conditioning. As generation proceeds, the model conditions on a longer history of its own outputs, increasing the opportunity for unsupported but locally plausible continuations and thus the relevance of drift-oriented hidden-state conflict. For shorter responses, the main challenge is more often to distinguish among plausible evidence-grounded candidates. Recall that T denotes the number of generated tokens. We define a length-dependent weighting function:

$$w(T) = \frac{1}{1+\exp(T - T_0)},$$

where T_0 is the transition length between short- and long-form regimes.

Final score. Finally, we define the overall hallucination score as

$$\mathcal{S}_{\text{final}} = (1 - w(T))\,\mathcal{S}_{\text{conflict}} - w(T)\,\mathcal{S}_{\text{likelihood}}. \tag{7}$$

Note that $\mathcal{S}_{\text{conflict}}$ (Eq. 6) is positively correlated with hallucination, while $\mathcal{S}_{\text{likelihood}}$ (Eq. 3) is negatively correlated. The subtraction in Eq. 7 aligns both into a single score where a higher $\mathcal{S}_{\text{final}}$ consistently indicates a greater likelihood of hallucination. Although the weighting emphasizes one signal more strongly in each regime, the two signals can provide complementary evidence in settings with mixed response lengths (Appendix B.4).

Remark. A practical advantage of the regime-aware design is that bidirectional likelihood only needs to be computed when it is expected to contribute meaningfully. For sufficiently long responses, where the detector is effectively governed by the hidden-state signal, the reasoning-generation step can be omitted as an inference-time optimization. As a result, the additional decoding cost of bidirectional likelihood is concentrated on short-form responses and bounded by fixed computing budgets; further analysis is provided in Section 4.3 and Appendix H.

4 Experiments

4.1 Settings

Datasets. For short-form evaluation, we use six open-book QA datasets from MrQA [5] (HotpotQA [30], TriviaQA [13], SearchQA [4], SQuAD [22], NewsQA [28], NQ [15]), where each question is paired with external context containing the correct answer. Under our assumption of accurate retrieval, we use these datasets as retrieval-grounded settings to study hallucination in the generation phase. An additional analysis on a perturbed NQ variant (NQ-swap) is provided in Appendix J. For long-form evaluation, we use two RAG hallucination benchmarks: RAGTruth [20], a manually annotated benchmark covering three task types (Question Answering, Data-to-Text Writing, and News Summarization), and Dolly (AC), a dataset from [9], where responses are assessed based on their consistency with external context. These benchmarks together span concise QA answers and longer descriptive generations; response-length statistics and per-dataset analysis are in Appendix B.4.

Baselines. We compare ARGUS against two categories of hallucination detection baselines. (1) *Model-output-based methods*: logit-based uncertainty estimators (Perplexity [23], LN-Entropy [18], Focus [33]), sampling-based consistency estimators (SelfCheckGPT [19]), and verbalization-based estimators (P(True) [14], RefChecker [9]). (2) *Internal-representation-based methods*: EigenScore [2], LLMCheck [25], ReDeEP [26], and LUMINA [31].

Metrics. We evaluate detection performance using AUROC (AUC), which measures the trade-off between true and false positive rates, and AUPRC (PRC), which captures the precision-recall trade-off. Both are threshold-independent. We further report F_1, where the decision threshold for each method is determined on a common random subset of the training data.

Implementation details. We conduct experiments on three backbones: LLaMA2-7B, LLaMA2-13B [27], and LLaMA3-8B [8], using the base version for short-form evaluation and the instruction-tuned variants for long-form evaluation. Results on an additional backbone, Mistral-7B [12], are reported in Appendix K. For the bidirectional likelihood, we set $K = 3$ and generate the reasoning sequence $\mathbf{z}$ via greedy decoding with a budget of 20 tokens. The transition length (T_0) is set to 10. All ARGUS hyperparameters are fixed across datasets and backbones, and all baselines are reproduced under our setup. Further details are in Appendix A.

4.2 Main Results

Table 1 reports detection performance on short-form benchmarks (six open-book QA datasets) and long-form benchmarks (RAGTruth and Dolly); per-dataset results are provided in Appendix N. Across all three backbones, ARGUS achieves strong overall performance across regimes: by adaptively combining regime-specialized signals, ARGUS achieves the best performance in most short- and long-form settings, while prior methods tend to specialize in one regime.

Short-form regime. ARGUS substantially improves over Perplexity in AUC consistently across all three backbones (0.7392 to 0.8077 on LLaMA2-7B, 0.7630 to 0.8184 on LLaMA2-13B, 0.6700

Model	Method	Short-form (6 QA datasets)			Long-form (RAGTruth, Dolly)		
		AUC	PRC	F_1	AUC	PRC	F_1
LLaMA2-7B	Perplexity [23]	0.7392	0.7552	0.7542	0.4140	0.4842	0.7053
	LN-Entropy [18]	0.7029	0.7292	0.7394	0.6832	0.6974	0.7134
	Focus [33]	0.5629	0.6673	0.7129	0.6009	0.6521	0.7046
	SelfCKGPT [19]	0.6759	0.7033	0.7273	0.6071	0.6374	0.7169
	EigenScore [2]	0.6933	0.6934	0.7358	0.5863	0.6229	0.7195
	P(True) [14]	0.6611	0.7156	0.7089	0.6355	0.6572	0.7069
	RefChecker [9]	0.5314	0.5821	0.7132	0.5764	0.6079	0.7038
	LLMCheck [25]	0.5083	0.5691	0.7130	0.6927	0.6845	0.7245
	ReDeEP [26]	0.6629	0.7124	0.6185	**0.7058**	0.7218	0.7141
	LUMINA [31]	0.5696	0.6205	0.7141	0.6455	0.6900	0.7142
	ARGUS (ours)	**0.8077**	**0.8139**	**0.7850**	0.6894	**0.7419**	**0.7317**
LLaMA2-13B	Perplexity [23]	0.7630	0.7299	0.7236	0.3823	0.4272	0.6743
	LN-Entropy [18]	0.7298	0.7024	0.7096	0.6939	0.6812	0.7098
	Focus [33]	0.5837	0.6314	0.6566	0.6922	0.6780	0.6906
	SelfCKGPT [19]	0.6999	0.6875	0.6937	0.6186	0.6312	0.7026
	EigenScore [2]	0.7107	0.6620	0.7080	0.6533	0.6192	0.7015
	P(True) [14]	0.3848	0.4400	0.6625	0.6909	0.7020	0.6906
	RefChecker [9]	0.5241	0.5235	0.6641	0.6212	0.5812	0.6680
	LLMCheck [25]	0.5189	0.5233	0.6638	0.7269	0.6900	0.7201
	ReDeEP [26]	0.5827	0.5755	0.6624	0.7616	0.7687	0.7442
	LUMINA [31]	0.5782	0.5746	0.6641	0.7286	0.7332	0.7378
	ARGUS (ours)	**0.8184**	**0.7947**	**0.7562**	**0.7999**	**0.7892**	**0.7778**
LLaMA3-8B	Perplexity [23]	0.6700	0.7363	0.7650	0.6647	0.6678	0.6656
	LN-Entropy [18]	0.6465	0.7245	0.7576	0.6766	0.6808	0.6587
	Focus [33]	0.5753	0.6965	0.7403	0.6194	0.5820	0.6628
	SelfCKGPT [19]	0.6605	0.7422	0.7606	0.5944	0.4835	0.6649
	EigenScore [2]	0.6856	0.7176	0.7650	0.6007	0.5778	0.6692
	P(True) [14]	0.6625	0.7441	0.7404	0.6519	0.6645	**0.7103**
	RefChecker [9]	0.5368	0.6310	0.7449	0.5740	0.5340	0.6435
	LLMCheck [25]	0.5040	0.6099	0.7427	0.6798	0.6573	0.6896
	ReDeEP [26]	0.6405	0.7287	0.7331	0.6968	0.6813	0.6826
	LUMINA [31]	0.5552	0.6617	0.7428	0.7235	0.7298	0.6837
	ARGUS (ours)	**0.7271**	**0.7785**	**0.7768**	**0.7279**	**0.7510**	0.6991

Table 1: Results on hallucination detection, averaged over 6 QA datasets for the short-form setting and over RAGTruth and Dolly for the long-form setting. Our method achieves strong overall performance across both regimes. Best results are boldfaced, and second-best results are underlined.

to 0.7271 on LLaMA3-8B). ARGUS outperforms other output-based methods (e.g., LN-Entropy, P(True)) by larger margins, driven by the bidirectional likelihood signal that discriminates among context-relevant alternatives via a reasoning-conditioned consistency check. In contrast, conflict-based methods (ReDeEP, LUMINA) are markedly weaker in this regime, consistent with the view that context–knowledge conflict signals are less informative for candidate misselection.

Long-form regime. On LLaMA2-13B, ARGUS attains AUC 0.7999, notably outperforming ReDeEP (0.7616) and LUMINA (0.7286), as well as output-based methods like LN-Entropy (0.6939); on LLaMA3-8B, ARGUS leads all baselines including LUMINA; on LLaMA2-7B, ARGUS is slightly below ReDeEP (0.6894 vs 0.7058), with clear leads on LLaMA2-13B and LLaMA3-8B. These results suggest that our hidden-state conflict score effectively captures the gradual long-form drift missed by output-based methods, demonstrating consistently strong performance across backbones with leads over prior conflict-based baselines in most settings.

Type	Method	Short	Long
Model Output	Perplexity [23]	0.7630	0.3823
Internal	ReDeEP [26]	0.5827	0.7616
	LUMINA [31]	0.5782	0.7286
Ours	$\mathcal{S}_{\text{likelihood}}$	0.8015	0.3575
	$\mathcal{S}_{\text{conflict}}$	0.5229	0.7999
	$\mathcal{S}_{\text{final}}$ **(ARGUS)**	**0.8184**	**0.7999**

Table 2: Effect of each component.

Type	Method	Short	Long
Model Output	Perplexity [23]	0.7630	0.3823
Internal	ReDeEP [26]	0.5827	0.7616
	LUMINA [31]	0.5782	0.7286
Combined	ReDeEP + Perplexity (fixed sum)	0.7420	0.4708
	LUMINA + Perplexity (fixed sum)	0.7335	0.4889
	$\mathcal{S}_{\text{likelihood}}$ + $\mathcal{S}_{\text{conflict}}$ (fixed sum)	0.7238	0.6796
	ARGUS (ours, adaptive)	**0.8184**	**0.7999**

Table 3: Naive combinations vs. adaptive ARGUS.

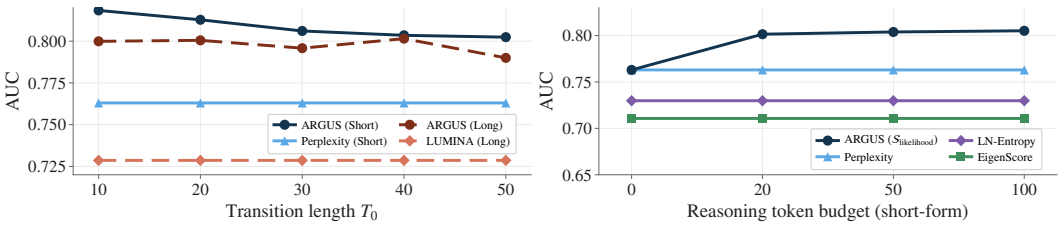


Figure 5: (Left) AUC across transition length T_0. (Right) AUC across reasoning token budget.

4.3 Ablation Study

Effect of Each Component. Table 2 reports AUC for each component on LLaMA2-13B alongside representative baselines. In the short-form regime, $\mathcal{S}_{\text{likelihood}}$ substantially improves over Perplexity (0.7630 to 0.8015); since Perplexity corresponds to the forward likelihood (Eq. 1) as a detection signal, this gap directly isolates the contribution of the bidirectional consistency check (further decomposition with qualitative examples in Appendix E). In the long-form regime, $\mathcal{S}_{\text{conflict}}$ outperforms ReDeEP (0.7616) and LUMINA (0.7286), showing that hidden-state decomposition provides a more discriminative signal of context–knowledge conflict than these baselines.

Each component specializes in its target regime: $\mathcal{S}_{\text{likelihood}}$ excels in short-form but drops to 0.3575 in long-form, while $\mathcal{S}_{\text{conflict}}$ excels in long-form but drops to 0.5229 in short-form. This regime-specific behavior motivates adaptive weighting: $\mathcal{S}_{\text{final}}$ improves over $\mathcal{S}_{\text{likelihood}}$ in short-form (0.8015 to 0.8184), suggesting that $\mathcal{S}_{\text{likelihood}}$ and $\mathcal{S}_{\text{conflict}}$ are complementary in this regime (see Appendix B.4 for per-dataset analysis), and matches $\mathcal{S}_{\text{conflict}}$'s long-form AUC (0.7999), since adaptive weighting downweights $\mathcal{S}_{\text{likelihood}}$ at long response lengths.

Why adaptive weighting? A natural alternative to adaptive weighting is to fuse output-based and context–knowledge conflict signals via a fixed combination rule (e.g., summation). Table 3 compares the AUC of ARGUS against such combinations of Perplexity with ReDeEP and LUMINA on LLaMA2-13B. In short-form, both combinations fall below Perplexity alone. In long-form, both combinations degrade substantially below ReDeEP and LUMINA alone. The two regimes exhibit different dominant failure patterns (candidate misselection vs. gradual drift): output-likelihood and conflict signals differ in their relative effectiveness across regimes, so simple combination dilutes the more informative signal with the less informative one. A fixed sum of our two proposed signals likewise underperforms the stronger of the two in each regime. ARGUS sidesteps this by adaptive weighting, outperforming these simple combinations in both regimes. Applying the same length-adaptive weighting to the prior signal pairs improves them as well, showing that the benefit of the rule is not tied to our signals, with our signal pair achieving the highest AUC in both regimes under the same adaptive weighting (Appendix C).

Robustness to Transition Length T_0. ARGUS uses a transition length T_0 to weight short-form and long-form signals. As in Figure 5 (left), AUC varies mildly across a wide range of T_0, while consistently outperforming Perplexity in the short-form regime and LUMINA in the long-form regime. This robustness supports response length as a practical proxy without careful tuning.

Effect of Reasoning Budget. We vary the reasoning budget for $\mathcal{S}_{\text{likelihood}}$ to examine how additional reasoning affects detection performance. When the budget is 0, $\mathcal{S}_{\text{likelihood}}$ reduces to standard forward likelihood (i.e., Perplexity). As shown in Figure 5 (right) on LLaMA2-13B, short-form AUC improves substantially as the budget increases from 0 to 20, and then largely saturates, indicating that

most of the benefit is obtained with a moderate reasoning budget; we adopt 20 tokens by default. We focus on short-form settings here, where the reasoning budget directly affects $\mathcal{S}_{\text{likelihood}}$.

Additional experiments and analyses. Additional analyses, including detailed analyses of our $\mathcal{S}_{\text{likelihood}}$ and $\mathcal{S}_{\text{conflict}}$ components, complexity analysis, and full per-dataset results, are in Appendix.

5 Conclusion

We provided evidence that different detection signals are informative in different response regimes. In short-form tasks, hallucination typically manifests as selecting an incorrect answer among plausible candidates, which our bidirectional likelihood score is designed to detect by evaluating consistency between the response and a generated reasoning step. In long-form generation, hallucination may emerge as gradual drift from the evidence, which our hidden-state conflict measure is designed to capture by comparing how each generated token aligns with the retrieved context versus the model's internal knowledge. By adaptively combining these signals based on response length, ARGUS achieves strong overall performance across 8 diverse short- and long-form datasets. We believe regime-aware detection is a promising direction for improving the reliability of RAG systems.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2024-00340966, No. RS-2024-00408003, and No. RS-2024-00345351); by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2024-00444862 and No. RS-2025-02283048, Developing the Next-Generation General AI with Reliability, Ethics, and Adaptability); by the IITP under the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25544647) grant funded by the Korea government (MSIT); and by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (MSIT) (No. S26120000). Jungwuk Park is the corresponding author.

References

[1] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

[2] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

[3] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

[4] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

[5] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd workshop on machine reading for question answering*, pages 1–13, 2019.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

[7] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*, 2024.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[13] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

[14] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[15] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[17] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063, 2021.

[18] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

[19] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023.

[20] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, 2024.

[21] nostalgebraist. Interpreting GPT: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, 2020.

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392, 2016.

[23] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2022.

[24] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, 2021.

[25] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

[26] Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024.

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[28] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, 2017.

[29] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, 2024.

[30] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380, 2018.

[31] Samuel Yeh, Sharon Li, and Tanwi Mallick. Lumina: Detecting hallucinations in rag system with context-knowledge signals. *arXiv preprint arXiv:2509.21875*, 2025.

[32] Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *arXiv preprint arXiv:2309.15098*, 2023.

[33] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, 2023.

Appendix

A Implementation Details

This appendix provides additional implementation details for ARGUS and the baselines.

Backbone models. We use three backbone language models: LLaMA2-7B [27], LLaMA2-13B [27], and LLaMA3-8B [8]. For each backbone, we use the base version for short-form evaluation and the instruction-tuned variant (chat or instruct) for long-form evaluation.

ARGUS hyperparameters. For the bidirectional likelihood, we set $K = 3$ and a maximum budge of 20 tokens for the reasoning sequence $\mathbf{z}$. All generation in our method is performed via greedy decoding. For the conflict score $\mathcal{S}_{\text{conflict}}$, we aggregate contributions across the bottom 32 transformer layers, applied uniformly across all backbones. The transition length for adaptive weighting is $T_0 = 10$. All hyperparameters are fixed uniformly across datasets and backbones.

Baselines. For all baseline methods, we use hyperparameters as specified in their official repositories or as reported in the original papers.

Threshold selection. Since our method requires no training, the decision threshold for F_1 evaluation is selected on a small subset of the training data. Specifically, we sample 1% of the training set for short-form datasets and 20% for long-form datasets, using random seed 0. The threshold that maximizes F_1 on this subset is then applied for evaluation. The same procedure is applied to all baseline methods for a fair comparison.

Hardware. All reported metrics were obtained on a 4×NVIDIA RTX 3090 and an NVIDIA RTX PRO 6000 setup. Inference-time measurements were conducted on a 3×NVIDIA RTX 3090 setup.

B Datasets and Per-Dataset Analysis

B.1 Short-form Datasets

Multi-Hop Reasoning (HotpotQA). HotpotQA is a benchmark dataset for multi-hop reasoning, requiring models to synthesize information across multiple paragraphs to generate an answer. This dataset emphasizes complex reasoning over distributed evidence, making it a critical benchmark for evaluating multi-hop question answering.

Retrieval-Based QA (TriviaQA, SearchQA). TriviaQA and SearchQA require reasoning over longer contexts, with answers scattered across retrieved documents. These datasets test the model's ability to focus on relevant content in lengthy contexts and generate precise answers.

Single-Paragraph Extraction (SQuAD, NewsQA). SQuAD and NewsQA are standard extractive QA datasets where the answer is typically located within a single paragraph. These datasets are widely used for evaluating the span-extraction capabilities of QA systems.

Document-Level QA (NQ). NQ is a document-level open-domain question answering dataset driven by real user queries. It requires systems to extract long answers from entire Wikipedia documents and generate specific short answers, evaluating document-level information retrieval and natural language understanding capabilities.

B.2 Long-form Datasets

RAGTruth. The RAGTruth dataset [20] is a high-quality, manually annotated RAG hallucination benchmark covering three task types: Question Answering, Data-to-Text Writing, and News Summarization. Each data point consists of a query, retrieved documents, an LLM-generated answer, and span-level hallucination annotations. RAGTruth provides answers generated by six different LLMs, including Llama2-7B, Llama2-13B, Llama2-70B, Mistral-7B, GPT-3.5, and GPT-4. In our experiment, we also utilize the extended test set provided by [26], who curated and annotated Llama3-8B generated responses.

Dolly (AC). The Dolly (AC) dataset is sourced from [9] and consists of tasks such as text summarization, closed-QA, and information extraction. Responses are assessed based on their consistency with

	Short							Long	
Model	HotpotQA	TriviaQA	SearchQA	SQuAD	NewsQA	NQ	NQ-swap	RAGTruth	Dolly
LLaMA2-7B	5.02	3.57	3.27	6.36	6.80	7.57	4.41	204.98	90.88
LLaMA2-13B	5.02	3.57	3.26	6.36	6.79	7.57	4.41	207.78	81.52
LLaMA3-8B	3.92	3.15	2.90	5.12	5.42	5.99	3.45	149.16	42.19

Table 4: Average response length (in tokens) per dataset for each backbone model.

Backbone	Quantity	HotpotQA	TriviaQA	SearchQA	SQuAD	NewsQA	NQ
LLaMA2-7B	$T \geq T_0$	8.3%	0.8%	0.6%	**17.1%**	**20.5%**	**22.2%**
	$\mathcal{S}_{\text{likelihood}}$	0.8177	0.7561	0.8145	0.7759	0.7868	0.7935
	$\mathcal{S}_{\text{conflict}}$	0.4781	0.3822	0.3330	0.4202	0.4974	0.4910
	$\mathcal{S}_{\text{final}}$	0.8181	0.7628	0.8158	**0.7888**	**0.8101**	**0.8500**
LLaMA2-13B	$T \geq T_0$	8.3%	0.7%	0.6%	**17.1%**	**20.4%**	**22.2%**
	$\mathcal{S}_{\text{likelihood}}$	0.8312	0.7362	0.8032	0.8097	0.8472	0.7815
	$\mathcal{S}_{\text{conflict}}$	0.5221	0.4409	0.3763	0.5477	0.6239	0.6266
	$\mathcal{S}_{\text{final}}$	0.8279	0.7430	0.8044	**0.8258**	**0.8701**	**0.8392**
LLaMA3-8B	$T \geq T_0$	2.8%	0.3%	0.2%	**11.4%**	**13.5%**	**14.8%**
	$\mathcal{S}_{\text{likelihood}}$	0.7182	0.7270	0.7452	0.6025	0.6372	0.7565
	$\mathcal{S}_{\text{conflict}}$	0.3574	0.4409	0.3903	0.3468	0.5058	0.4647
	$\mathcal{S}_{\text{final}}$	0.7263	0.7383	0.7497	**0.6494**	**0.6957**	**0.8029**

Table 5: Per-dataset $T \geq T_0$ proportion and component scores.

the provided external context, focusing on the alignment between the generated responses and the retrieved evidence.

B.3 Hallucination Labels

For short-form retrieval-grounded QA, we operationalize hallucination as the model selecting an answer that does not match the gold answer under accurate context, rather than annotating individual unfaithful spans within the response. Specifically, a response is labeled as hallucinated if the generated answer does not exactly match any of the ground-truth answers under SQuAD-style normalization (lowercasing, removal of punctuation and articles, and whitespace collapse). When the dataset provides multiple acceptable gold answers, the response is labeled as faithful if it matches any of them. For long-form datasets, we use the provided annotations in both cases. For RAGTruth (excluding Llama3-8B responses), a response is labeled as hallucinated if it contains at least one hallucinated span according to the provided span-level annotations. For RAGTruth (Llama3-8B) and Dolly (AC), we directly use the provided response-level hallucination labels.

B.4 Per-Dataset Length Statistics and Component Analysis

Average response length per dataset. Table 4 reports the average response length in tokens for each dataset and backbone model. Short-form datasets exhibit consistently short responses, ranging from 2.9 to 7.6 tokens on average, whereas long-form datasets show substantially longer responses, with RAGTruth averaging 149–207 tokens and Dolly averaging 42–90 tokens. This length pattern is consistent across all three backbone models.

$\mathcal{S}_{\text{conflict}}$ **contributes on long-tail responses within QA benchmarks.** We further break down the short-form gain of $\mathcal{S}_{\text{final}}$ over $\mathcal{S}_{\text{likelihood}}$ (Table 2: 0.8184 vs. 0.8015) per dataset. Table 5 reports, for each short-form QA dataset and backbone, the proportion of responses with length $T \geq T_0$ alongside the AUC of $\mathcal{S}_{\text{likelihood}}$, $\mathcal{S}_{\text{conflict}}$, and the adaptively combined $\mathcal{S}_{\text{final}}$. While TriviaQA and SearchQA are almost entirely in the short regime (under 1% of responses exceed T_0), SQuAD, NewsQA, and NQ contain a non-trivial long-tail (11–22% across the three backbones). On these datasets, $\mathcal{S}_{\text{final}}$ shows clear improvements over $\mathcal{S}_{\text{likelihood}}$ (up to +0.058 AUC on NQ for LLaMA2-13B) by incorporating $\mathcal{S}_{\text{conflict}}$ on the long-tail subset, while datasets without a meaningful long-tail show only marginal differences. This suggests that adaptive weighting operates at the response level, not the dataset level:

$\mathcal{S}_{\text{conflict}}$ receives substantially greater weight for individual responses that exceed T_0, even within QA benchmarks.

No dataset metadata at inference time. The adaptive weighting is a function of realized response length only, with T_0 fixed across all datasets and backbones. No dataset identity or task type enters the weighting computation; instead, response length serves as a practical proxy in the weighting rule. This allows the same weighting rule to be applied across short- and long-form retrieval-grounded settings without per-dataset tuning.

Effect of per-signal normalization. $\mathcal{S}_{\text{conflict}}$ (a difference of cosine similarities) and $\mathcal{S}_{\text{likelihood}}$ (a product of likelihoods) lie on different scales. We assess the effect of this scale mismatch on aggregate detection performance by applying per-signal z-normalization before combination, with the mean and standard deviation of each signal estimated per dataset on the subset used for F_1 threshold selection and applied to the test set. Because $w(T)$ is near-saturated for most responses, $\mathcal{S}_{\text{final}}$ is typically dominated by one signal, and normalization yields nearly identical aggregate results on LLaMA2-13B: short-form macro AUC 0.8179 with normalization vs 0.8184 without, and long-form 0.7999 vs 0.7999, with a maximum per-dataset change of 0.006. We therefore keep the uncalibrated formulation, which avoids an additional dependence on calibration statistics.

Behavior within the transition band. On the full short-form set, the combined score reaches 0.8184 macro AUC against 0.8015 for $\mathcal{S}_{\text{likelihood}}$ alone (Table 2). Within the band $7 \leq T \leq 13$ around $T_0 = 10$ (9,047 short-form responses on LLaMA2-13B), where $w(T)$ moves from 0.95 to 0.05, however, $\mathcal{S}_{\text{likelihood}}$ alone reaches 0.8376 macro AUC, above the combined score (0.7540, or 0.8118 with normalization), while $\mathcal{S}_{\text{conflict}}$ alone is near chance (0.4950). The weak conflict signal in this band may reflect limited opportunity for drift to accumulate in short responses (Appendix F.1). Although the combination reduces discrimination within this band, ARGUS receives no task or regime labels, and its weighting rule depends only on the realized response length, so we retain a single fixed T_0 across datasets and backbones rather than adapting the transition to each task. The reported aggregate gains support this design, while the band-level results indicate room to improve the transition rule.

C Weighting Rule versus Signal Choice

Table 3 in the main text compares fixed-sum combinations of three signal pairs with adaptive ARGUS. Table 6 separates the two factors involved, which signals are combined and how they are weighted, by evaluating each of three signal pairs under both a fixed sum and the length-adaptive weighting $w(T)$ of Eq. 7 on LLaMA2-13B.

Combined signals	**Fixed sum**		**Adaptive** $w(T)$	
	Short	Long	Short	Long
Perplexity + ReDeEP	0.7420	0.4708	0.8093	0.7616
Perplexity + LUMINA	0.7335	0.4889	0.8080	0.7286
$\mathcal{S}_{\text{likelihood}}$ + $\mathcal{S}_{\text{conflict}}$ (ours)	0.7238	0.6796	**0.8184**	**0.7999**

Table 6: Fixed-sum versus length-adaptive combination for three signal pairs.

Two readings follow. On the weighting axis, the fixed sum underperforms the adaptive rule in both regimes for all three pairs, including ours; across the three pairs, the adaptive rule adds +0.067 to +0.095 AUC on short-form and +0.120 to +0.291 on long-form. This is consistent with each signal being primarily informative in a different regime, so that a fixed sum dilutes the more informative signal with the less informative one. That adaptive weighting also improves the prior pairs is expected, since the rule is signal-agnostic and operationalizes the regime finding of this paper. On the signal axis, under the same adaptive weighting, our signal pair achieves the highest AUC in both regimes among the tested pairs, with modest but consistent margins on short-form (+0.009 over the ReDeEP-based and +0.010 over the LUMINA-based pair) and clearer margins on long-form (+0.038 and +0.071); the long-form margins coincide with the component-level differences in Table 2, since $w(T)$ is near zero for long responses.

D Grounding Analysis of Short-form Hallucinations

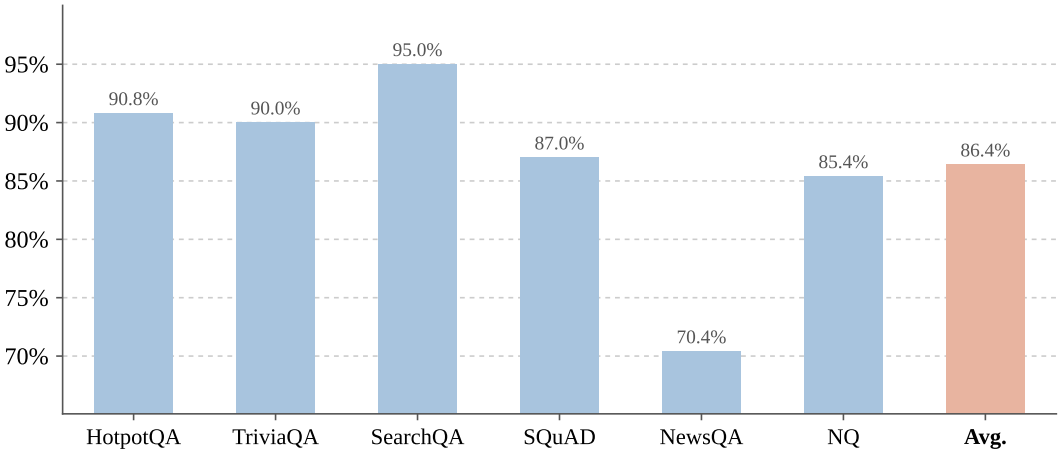


Figure 6: Proportion of grounded hallucinated responses across short-form QA datasets. The majority of hallucinated responses remain anchored in the provided input, suggesting that short-form hallucination predominantly arises from selecting an incorrect answer present in the context.

We investigate to what extent hallucinated responses in short-form settings remain tied to the provided context or question, rather than arising entirely from information absent from the input. Figure 6 shows, for each short-form QA dataset, the proportion of hallucinated responses (produced by greedy decoding with LLaMA2-7B) that still match the retrieved context or question under a normalized lexical criterion based on substring matching and ROUGE-L. On average, approximately 86% of hallucinated responses are matched to the provided input, providing suggestive empirical support that many short-form errors remain evidence-related. This helps explain why context–knowledge conflict signals are less discriminative in the short-form regime.

The remaining $\sim$14% should not be interpreted as a single error type. They may include genuinely unsupported answers generated from parametric knowledge, as well as paraphrastic or partial matches not captured by our lexical criterion. Grounding ratios are somewhat lower for NewsQA, which may reflect lexical matching missing paraphrastic answers.

E Detailed Analysis of $\mathcal{S}_{\text{likelihood}}$

E.1 Analysis of the Reasoning-Consistency Signal

We examine whether the reasoning $\mathbf{z}$ generated after the response provides a useful signal for distinguishing hallucinated from faithful responses. On LLaMA2-13B across the six short-form QA datasets, we use only the reasoning term $s(\mathbf{z} \mid \mathbf{q}, \mathbf{c}, \mathbf{x}^{(k)})$ of Eq. 2 to compare candidate responses, excluding their forward likelihoods, and measure how often an alternative candidate receives a higher reasoning likelihood than the produced response. We count this event on all hallucinated responses and, as a control, on all faithful responses (Table 7).

Response type	**Event measured using the reasoning term alone**	**Rate**
Hallucinated responses	Any alternative candidate scores higher than the produced wrong answer	39.1%
Faithful responses (control)	Any alternative candidate scores higher than the produced correct answer	30.9%

Table 7: Rate at which the reasoning term alone prefers an alternative candidate over the produced response.

The event occurs more often for hallucinated than for faithful responses (39.1% vs 30.9%), consistent with the intuition in Section 3.1. However, the substantial overlap shows that this event alone is

insufficient for reliable detection. In many hallucinated cases the reasoning stays with the produced wrong answer; these include the rationalization failure mode illustrated in Figure 7b, where the model generates plausible-sounding reasoning consistent with its wrong answer. This is why the forward-likelihood term is retained in $\mathcal{S}_{\text{likelihood}}$ (Table 8).

Decomposition of the AUC gain. On the same six datasets, $\mathcal{S}_{\text{likelihood}}$ improves the short-form AUC over the forward likelihood from 0.7630 to 0.8015 (Table 8). AUC measures, across all pairs of one hallucinated and one faithful response, how often the hallucinated one receives the higher risk score. Scoring the same pairs once with the forward likelihood and once with $\mathcal{S}_{\text{likelihood}}$, the latter newly orders 8.2% of the pairs correctly where the forward likelihood had them wrong, and mis-orders 4.4% that the forward likelihood had right; the difference between these two percentages equals the AUC gain exactly (+0.039 using unrounded values). Thus, $\mathcal{S}_{\text{likelihood}}$ corrects more misordered pairs than it introduces, yielding the observed net AUC gain.

E.2 Complementarity of Forward and Bidirectional Likelihoods: Quantitative Results

Model	Method	AUC	PRC	$\mathbf{F_1}$
LLaMA2-7B	Forward likelihood (Eq. 1)	0.7392	0.7552	0.7542
	Bidirectional likelihood (Eq. 2)	0.7368	0.7851	0.7202
	$\mathcal{S}_{\text{likelihood}}$ (Eq. 3)	0.7908	0.8028	0.7781
LLaMA2-13B	Forward likelihood (Eq. 1)	0.7630	0.7299	0.7236
	Bidirectional likelihood (Eq. 2)	0.7176	0.7383	0.6649
	$\mathcal{S}_{\text{likelihood}}$ (Eq. 3)	0.8015	0.7764	0.7417
LLaMA3-8B	Forward likelihood (Eq. 1)	0.6700	0.7363	0.7650
	Bidirectional likelihood (Eq. 2)	0.6491	0.7383	0.7445
	$\mathcal{S}_{\text{likelihood}}$ (Eq. 3)	0.6978	0.7557	0.7726

Table 8: Comparison of forward likelihood, bidirectional likelihood, and their combination $\mathcal{S}_{\text{likelihood}}$ on short-form QA datasets. Forward likelihood corresponds to the Perplexity baseline.

Table 8 decomposes $\mathcal{S}_{\text{likelihood}}$ into its two factors. The **forward likelihood** captures the model's standalone token-level confidence in the response, while the **bidirectional likelihood** captures the response's comparative standing among the top-K candidates after incorporating reasoning consistency. The combined score $\mathcal{S}_{\text{likelihood}}$ consistently outperforms both components alone across all three backbones, indicating that the two signals are complementary. Their product yields a stronger detection signal than either factor alone. Qualitative examples illustrating this complementarity are provided in Section E.3.

E.3 Complementarity of Forward and Bidirectional Likelihoods: Qualitative Examples

Figure 7 presents examples from HotpotQA and NQ that illustrate the complementary roles of the two components. In each pair, one signal remains informative while the other provides limited additional discrimination.

Forward likelihood is less informative, while bidirectional likelihood catches the error. In Figure 7a, the model produces a context-grounded but incorrect answer with high forward likelihood: in this multi-hop case, the model returns an intermediate entity (the school name *Lincoln Christ's*) instead of the final answer (the road *Wragby Road*). Forward likelihood remains high (0.76) because the response is locally fluent and textually supported by the context. By contrast, the bidirectional likelihood drops substantially (0.38) because the top-K candidates include the correct answer (*Wragby Road*), and this alternative receives comparable or stronger support under the reasoning-aware comparison.

Bidirectional likelihood provides limited additional discrimination, while forward likelihood remains informative. In Figure 7b, the bidirectional likelihood does not clearly separate the incorrect answer from plausible alternatives because the reasoning-conditioned comparison remains ambiguous. In this table-reasoning case, the model returns *Joe* instead of *Sal* (who has the highest punishment count) and fabricates a numerically plausible-sounding reasoning trace, so the reasoning-aware

Query: Agnes Body was the founding head of a state secondary school located on what road?

Retrieved Context:

Agnes Body: Agnes Body (29 April, 1866 – 31 March, 1952) was a British headmistress. She was the founding head of Lincoln Christ's Hospital Girls' High School and Queen Margaret's School, then in Scarborough. Lincoln Christ's Hospital School: Lincoln Christ's Hospital School is a state secondary school with academy status located on "Wragby Road" in Lincoln, Lincolnshire, England.

Model Response: Lincoln Christ's

Correct Answer: Wragby Road

Reasoning: Lincoln Christ's Hospital School is a state secondary school with academy status located on "Wragby Road" in Lincoln, Lincolnshire, England.

***K-1* Candidate Responses:** Wragby Road, Agnes Body was

Likelihood: 0.7614 **Bidirectional Likelihood:** 0.3804

(a) Likelihood fails to detect, bidirectional detects.

Query: who gets the most punishments on impractical jokers

Retrieved Context:

Impractical Jokers statistics

Joker	Punishments	Thumbs down	Thumbs up	Challenges
Joe	36 (23.4%)	136 (35.9%)	254 (68.5%)	377 (92.0%)
Murr	50 (34.0%)	168 (48.7%)	228 (59.3%)	379 (92.6%)
Q	40 (26.1%)	140 (38.0%)	252 (69.0%)	375 (91.4%)
Sal	55 (36.0%)	163 (42.6%)	239 (66.2%)	377 (92.0%)

Model Response: Joe

Correct Answer: Sal

Reasoning: 377 punishments out of 400 punishments were given to Joe.

***K-1* Candidate Responses:** M, Q

Likelihood: 0.3913 **Bidirectional Likelihood:** 0.7026

(b) Likelihood detects, bidirectional fails to detect.

Figure 7: Examples from HotpotQA (a) and NQ (b).

comparison stays consistent with the wrong response. As a result, the bidirectional likelihood for the wrong answer remains relatively high (0.70). However, forward likelihood is low (0.39), reflecting lower model confidence and thereby serving as a complementary signal.

In summary, forward likelihood remains useful when the reasoning-aware comparison is insufficiently discriminative among plausible candidates, while bidirectional likelihood is particularly useful for confidently wrong answers that lose support under the reasoning-aware comparison. Their combination therefore helps reduce both failure patterns, consistent with the quantitative gains reported in Table 8.

E.4 Comparison with QA Re-ranking

In short-form open-book QA, hallucination labels are determined by normalized exact-match correctness, and $\tilde{s}$ compares the produced response against the top-K candidates. A natural question is therefore whether the short-form gain of $\mathcal{S}_{\text{likelihood}}$ can be explained by QA re-ranking over the candidate set. We note first that ARGUS does not use the gold answer or the hallucination label, and it assigns a risk score to the already produced response rather than selecting or replacing it; the candidates serve as a tractable reference set for the normalization in Eq. 2. We then examine the question empirically on LLaMA2-13B across the six short-form QA datasets.

Direct QA re-ranking baseline. We prompt the same model with the context, the question, and the top-K candidates presented as (A)/(B)/(C) choices, using the template `Given the following information: {context} Question: {question} Candidates: (A) {...} (B) {...} (C) {...} Which candidate is best supported by the information? Answer:`. The model is thus asked to select the best-supported candidate; we normalize the next-token probabilities of the three choice labels and use one minus the probability assigned to the produced response as its risk score. Candidate order is shuffled per sample with a fixed seed, so the produced response does not always occupy the same position. Table 9 compares this baseline with $\mathcal{S}_{\text{likelihood}}$ and with the answer re-verification baselines from Table 1.

$\mathcal{S}_{\text{likelihood}}$ exceeds the direct QA re-ranking baseline on all six datasets (macro +0.074 AUC), and the baseline trails even Perplexity. These results suggest that the improvement is not fully explained by direct QA re-ranking over the same candidates.

Evaluation when the gold answer is absent from the candidates. We restrict evaluation to all faithful responses and the hallucinated responses whose gold answer does not appear among the top-K candidates under normalized matching (30,340 faithful and 19,007 hallucinated responses). On this set, $\mathcal{S}_{\text{likelihood}}$ still outperforms Perplexity (0.7871 vs 0.7551 macro AUC), showing that the gain persists when hallucinated responses lack a gold-matching candidate. We also note that the $\sim$86% statistic in Appendix D concerns whether the produced wrong answer is anchored in the

Method	Short-form AUC
Perplexity	0.7630
SelfCheckGPT	0.6999
P(True)	0.3848
Direct QA re-ranking over the top-K candidates	0.7276
$\mathcal{S}_{\text{likelihood}}$ (ours)	**0.8015**

Table 9: Comparison with a direct QA re-ranking baseline and existing detectors on short-form QA.

provided input; it does not imply that the correct answer is available among the candidates, which holds for only 31.2% of hallucinated responses.

E.5 Sensitivity to the Reasoning Step

The reasoning $\mathbf{z}$ is self-generated and can rationalize a wrong answer (Figure 7b). We therefore vary how $\mathbf{z}$ is produced on LLaMA2-13B (HotpotQA and NQ), keeping the QA prompt, the candidates, and the scoring fixed: the reasoning instruction (two rewordings of the instruction in Appendix I that change both the phrasing and the answer cue), the decoding strategy (greedy vs temperature sampling at 0.7 with three seeds), and the reasoning model ($\mathbf{z}$ generated by Mistral-7B instead of the backbone). The instruction variants are:

- **Paper prompt:** `Refer to the given information and briefly state the reason behind your previous answer. Reason:`
- **Paraphrase 1:** `Based only on the given information, briefly explain why your previous answer is correct. Explanation:`
- **Paraphrase 2:** `Briefly justify your previous answer using only the given information. Justification:`

Reasoning variant	HotpotQA	NQ
Paper prompt, greedy	0.8279	0.8392
Paraphrase 1	0.8258	0.8427
Paraphrase 2	0.8289	0.8404
Sampling (temperature 0.7; mean $\pm$ std over 3 seeds)	0.8201 ± 0.0016	0.8343 ± 0.0006
Cross-model reasoning ($\mathbf{z}$ from Mistral-7B)	0.8340	0.8439

Table 10: Effect of the reasoning prompt, decoding strategy, and reasoning model on short-form detection.

Table 10 reports ARGUS AUC. All reported variant AUCs differ from the default by less than 0.01 on both datasets, suggesting that ARGUS performance is relatively stable under the tested changes to reasoning generation, consistent with the robustness to the reasoning budget (Figure 5, right) and to K (Appendix E.6).

E.6 Robustness to Top-K Candidates

The bidirectional likelihood compares the response against the top-K alternative candidates after incorporating reasoning consistency. We vary $K \in \{2, 3, 5\}$ on LLaMA2-13B with all other settings fixed ($T_0 = 10$, reasoning budget 20). Long-form metrics are unchanged across K because $\mathcal{S}_{\text{conflict}}$ does not depend on candidate selection.

Performance is robust across K: short-form AUC varies within 0.005 (0.8151–0.8200), with PRC and F_1 showing similarly small variations, and long-form metrics are exactly invariant. We adopt $K = 3$ as the default, balancing candidate diversity against the additional cost of generating reasoning over more candidates.

K	Short-form (6 QA datasets)			Long-form (RAGTruth, Dolly)		
	AUC	PRC	F_1	AUC	PRC	F_1
2	0.8151	0.7899	0.7549	0.7999	0.7892	0.7778
3 (default)	0.8184	0.7947	0.7562	0.7999	0.7892	0.7778
5	0.8200	0.7979	0.7560	0.7999	0.7892	0.7778

Table 11: Effect of top-K candidates.

F Detailed Analysis of $\mathcal{S}_{\text{conflict}}$

F.1 Conflict Drift in Long-Form Generation

Model	**Group**	**Front (0-30%)**	**Mid (30-70%)**	**Back (70-100%)**	Δ
LLaMA2-7B	Faithful	0.7360	0.7875	0.824	0.0880
	Hallucinated	0.7554	0.8346	0.8797	**0.1243**
LLaMA2-13B	Faithful	0.3249	0.4094	0.4462	0.1213
	Hallucinated	0.3712	0.4922	0.5492	**0.1780**
LLaMA3-8B	Faithful	0.7296	0.7900	0.7993	0.0697
	Hallucinated	0.7649	0.8347	0.8448	**0.0799**

Table 12: Mean conflict score $\mathcal{S}_{\text{conflict}}$ across positional segments (thirds) for faithful and hallucinated responses. Δ denotes the difference between the back (70-100%) and front (0-30%) segments. Larger Δ indicates a stronger drift toward parametric over context influence.

Section 3.2 introduced the hidden-state conflict score as a signal for long-form drift. Here, we provide a more detailed per-segment analysis and additional empirical support. Specifically, we partition each response into three positional segments: Front (0–30%), Mid (30–70%), and Back (70–100%). Table 12 reports the per-backbone mean $\mathcal{S}_{\text{conflict}}$ in each segment.

Across all three backbones, $\mathcal{S}_{\text{conflict}}$ increases from front to back for both faithful and hallucinated generations, suggesting that long-form generation becomes progressively more internally driven as generation proceeds. Crucially, the increase is consistently larger for hallucinated responses (Δ in the rightmost column), indicating a steeper drift away from context grounding. These results provide additional empirical support for the long-form drift hypothesis in Section 3.2.

F.2 Sensitivity to the Layer Range

$\mathcal{S}_{\text{conflict}}$ aggregates source-specific updates over the bottom 32 layers for all backbones (Appendix A). For LLaMA2-7B, LLaMA3-8B, and Mistral-7B, 32 is the full model depth, so no layers are excluded; LLaMA2-13B (40 layers) is the only backbone for which the cap truncates. Unlike the per-model selection of specific heads and layers in ReDeEP, no per-model choice is involved. Table 13 varies the aggregation range on LLaMA2-13B long-form, the only case in which the cap binds.

Aggregation range	**Long-form AUC**
Bottom 24 layers	0.8080
Bottom 32 layers (default)	0.7999
All 40 layers	0.7841
ReDeEP (Table 1)	0.7616

Table 13: Effect of the layer aggregation range on long-form detection.

Across the tested ranges, AUC ranges from 0.7841 to 0.8080, remaining above ReDeEP even when all 40 layers are included.

F.3 Component Ablation

Backbone	Method	AUC	PRC	F_1
LLaMA2-7B	Full $\mathcal{S}_{\mathrm{conflict}}$	0.6920	0.7429	0.7317
	w/o context	0.6982	0.7325	0.7322
	w/o query	0.6389	0.6995	0.7173
	w/o response	0.6850	0.7306	0.7398
	w/o ffn	0.6911	0.7327	0.7385
LLaMA2-13B	Full	0.7999	0.7892	0.7778
	w/o context	0.8045	0.7898	0.7761
	w/o query	0.7127	0.7302	0.7316
	w/o response	0.7739	0.7670	0.7369
	w/o ffn	0.7874	0.7710	0.7694
LLaMA3-8B	Full	0.7311	0.7517	0.6955
	w/o context	0.7030	0.7354	0.6787
	w/o query	0.7113	0.7158	0.6738
	w/o response	0.7266	0.7503	0.6914
	w/o ffn	0.7342	0.7533	0.6888
Average	**Full**	**0.7410**	**0.7613**	**0.7350**
	w/o context	0.7352	0.7526	0.7290
	w/o query	0.6877	0.7152	0.7076
	w/o response	0.7285	0.7493	0.7227
	w/o ffn	0.7376	0.7523	0.7322

Table 14: Component ablation of $\mathcal{S}_{\mathrm{conflict}}$ on long-form benchmarks (RAGTruth + Dolly), averaged over both datasets.

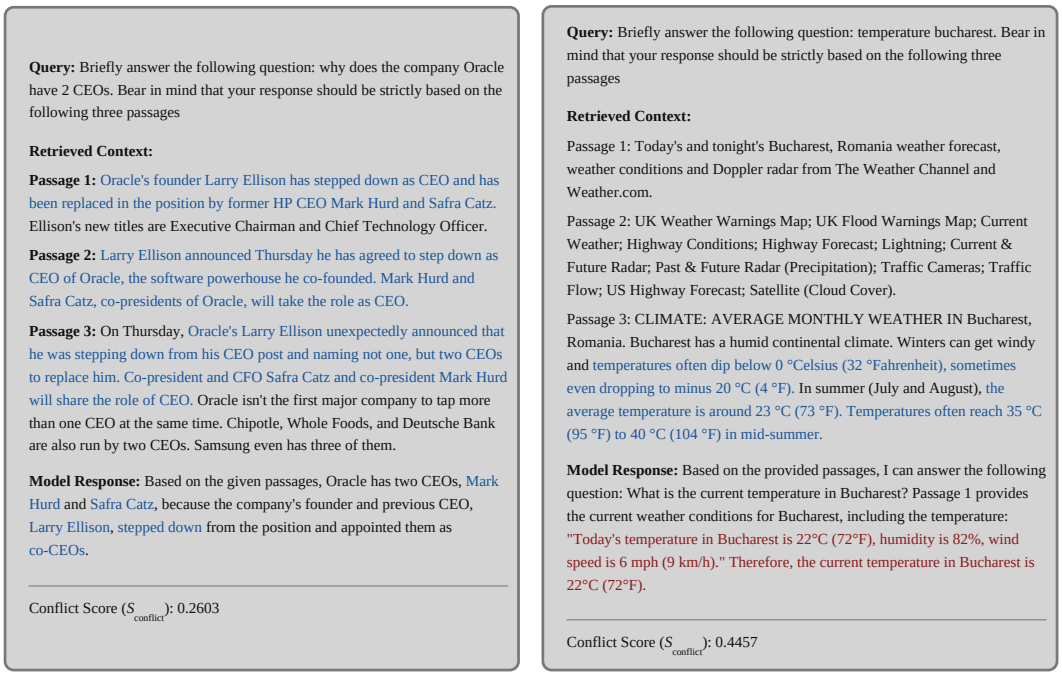
We ablate each of the four source-specific components (h_{context}, h_{query}, h_{response}, and h_{ffn}) by removing it from $\mathcal{S}_{\mathrm{conflict}}$ and re-evaluating on the long-form benchmarks. Across the three backbones, the full formulation achieves the best average AUC (0.741), PRC (0.7613), and F_1 (0.735). Removing the query component yields the largest drop (-0.053 average AUC), suggesting that query alignment provides particularly important information for long-form hallucination detection. Other removals lead to smaller changes (-0.003 to -0.013 AUC), indicating that the remaining components can partially compensate for a removed source. Thus, the source-decomposed score appears to degrade gracefully under single-component removal, rather than relying on a single component alone. Overall, while some ablated variants show small backbone-specific fluctuations, the full four-component formulation provides the most robust average performance across backbones.

F.4 Qualitative Examples of $\mathcal{S}_{\mathrm{conflict}}$

Figure 8 shows two long-form responses contrasting low and high $\mathcal{S}_{\mathrm{conflict}}$. In Figure 8a, the response is grounded in the retrieved context, with key entities (Mark Hurd, Safra Catz, Larry Ellison) traceable to the context, and $\mathcal{S}_{\mathrm{conflict}}$ remains low (0.2603). In Figure 8b, the response fabricates specific weather details (a temperature, humidity, and wind speed) that do not appear in the retrieved context, and $\mathcal{S}_{\mathrm{conflict}}$ is correspondingly higher (0.4457).

G Attention Decomposition: Detailed Analysis

In this section, we provide the detailed per-source decomposition $h_{\mathrm{s}}^{(t)}$ introduced in Section 3.2, covering the per-layer attention and FFN contributions, their accumulation in the residual stream, and the final-layer normalization. Throughout, $r^{(t)} \in \mathbb{R}^d$ denotes the residual stream just before the final-layer normalization, and $h^{(t)} = \mathrm{Norm}(r^{(t)}) \in \mathbb{R}^d$ is the final hidden state used for next-token prediction. For notational clarity, we omit the multi-head index throughout; the construction extends linearly across heads.



(a) Faithful response with low $\mathcal{S}_{\text{conflict}}$.

(b) Hallucinated response with high $\mathcal{S}_{\text{conflict}}$.

Figure 8: Long-form examples illustrating $\mathcal{S}_{\text{conflict}}$ discrimination. Tokens supported by the retrieved context are highlighted in blue, while tokens that contradict or are unsupported by the context are highlighted in red.

Attention output. In layer l, the attention output for token t is a weighted sum over all attended positions,

$$a^{(l,t)} = \left(\sum_i \alpha_i^{(l,t)} v_i^{(l,t)} \right) W_O^{(l)},$$

with $\alpha_i^{(l,t)} \in \mathbb{R}$ the attention weight at position i, $v_i^{(l,t)} \in \mathbb{R}^{d_v}$ the value vector, and $W_O^{(l)} \in \mathbb{R}^{d_v \times d}$ the output projection. Restricting the sum to each source set $\mathcal{P}_{\mathrm{s}} \in \{\mathcal{C}, \mathcal{Q}, \mathcal{R}_{<t}\}$ gives the per-source per-layer attention contribution:

$$h_{\mathrm{s}}^{(l,t)} = \left(\sum_{i \in \mathcal{P}_{\mathrm{s}}} \alpha_i^{(l,t)} v_i^{(l,t)} \right) W_O^{(l)},$$

for $\mathrm{s} \in \{\text{context}, \text{query}, \text{response}\}$, with $a^{(l,t)} = \sum_{\mathrm{s}} h_{\mathrm{s}}^{(l,t)}$ by linearity.

FFN contribution. The FFN contribution at layer l is $h_{\text{ffn}}^{(l,t)} = \mathrm{FFN}^{(l)}(\tilde{r}^{(l,t)})$, where $\tilde{r}^{(l,t)}$ is the input after the layer's pre-FFN normalization.

Layer updates and residual stream accumulation. In a pre-norm transformer (where layer normalization is applied to the input of each sublayer rather than after the residual addition), each layer applies attention and FFN with residual updates:

$$r^{(l+1/2,\,t)} = r^{(l,\,t)} + a^{(l,t)}, \qquad r^{(l+1,\,t)} = r^{(l+1/2,\,t)} + h_{\text{ffn}}^{(l,t)},$$

where $a^{(l,t)} = \mathrm{Attn}^{(l)}\big(\mathrm{LN}_{\text{attn}}^{(l)}(r^{(l,\,t)})\big)$ and $h_{\text{ffn}}^{(l,t)} = \mathrm{FFN}^{(l)}\big(\mathrm{LN}_{\text{ffn}}^{(l)}(r^{(l+1/2,\,t)})\big)$. Although $a^{(l,t)}$ and $h_{\text{ffn}}^{(l,t)}$ are computed through layer-internal nonlinearities, they enter the residual stream additively. Unrolling these updates across layers, the residual stream just before the final-layer normalization accumulates per-layer updates as

$$r^{(t)} = r_{\text{embed}}^{(t)} + \sum_{l=1}^{L} \big(a^{(l,t)} + h_{\text{ffn}}^{(l,t)}\big) = r_{\text{embed}}^{(t)} + \sum_{\mathrm{s}} \sum_{l=1}^{L} h_{\mathrm{s}}^{(l,t)},$$

where $r^{(t)}_{\text{embed}} = r^{(1,t)}$ is the input to the first transformer layer (the token embedding, with positional embedding added in models that use absolute or learned positional encodings). The second equality uses $a^{(l,t)} = \sum_{\text{s}} h^{(l,t)}_{\text{s}}$ and the index s ranges over $\{\text{context}, \text{query}, \text{response}, \text{ffn}\}$ (with $h^{(l,t)}_{\text{ffn}}$ taking the role of $h^{(l,t)}_{\text{s}}$ for $\text{s} = \text{ffn}$). Each $\sum_{l=1}^{L} h^{(l,t)}_{\text{s}}$ is the source-specific accumulated update written into the residual stream across these L layers.

Final-layer normalization. Rather than recomputing normalization statistics for each per-source partial sum, we use the shared final-layer normalization statistics of $r^{(t)}$. Our backbones (LLaMA and Mistral) use RMSNorm, $\text{RMSNorm}(x) = \gamma \odot x/\text{RMS}(x)$, with no bias and γ a learned per-element scale. We apply this with the RMS scale fixed to $\text{RMS}(r^{(t)})$ for all per-source partial sums (rather than recomputed per sum), giving $\text{Norm}(x) = \gamma \odot x/\text{RMS}(r^{(t)})$ in our construction. Concretely, for each $\text{s} \in \{\text{context}, \text{query}, \text{response}, \text{ffn}\}$, the source-wise vector is defined as

$$h^{(t)}_{\text{s}} = \text{Norm}\left(\sum_{l=1}^{L} h^{(l,t)}_{\text{s}}\right), \qquad h^{(t)}_{\text{embed}} = \text{Norm}\left(r^{(t)}_{\text{embed}}\right).$$

This maps each per-source component into the same normalized space as $h^{(t)}$, providing a comparable proxy in this shared space.

Additive decomposition and its interpretation. This yields the per-source decomposition $h^{(t)} = h^{(t)}_{\text{embed}} + \sum_{\text{s}} h^{(t)}_{\text{s}}$ in the normalized space, lifted from the additive structure of the residual stream. Because the shared RMS scale reduces normalization to a fixed element-wise linear rescaling, this additive identity carries over from the residual stream into the normalized space. The resulting per-source components are not causally independent: each per-layer attention output $h^{(l,t)}_{\text{s}}$ is computed from the layer input $r^{(l,t)}$, which already contains contributions from all sources at earlier layers. Thus, the decomposition should be interpreted as tracking where source-specific updates are written into the realized residual stream, rather than as a counterfactual statement about what would happen if a source were removed. Accordingly, we use $h^{(t)}_{\text{s}}$ as an interpretive indicator of each source's relative influence on the final hidden state, not as an independent causal effect.

H Complexity Analysis

We measure per-sample inference time on LLaMA2-13B with RTX 3090 GPUs, using a 20-token reasoning budget and $K = 3$ candidates. Table 15 reports this cost alongside detection AUC (Section 4.2) for ARGUS when each component is computed alone ($\mathcal{S}_{\text{conflict}}$ only / $\mathcal{S}_{\text{likelihood}}$ only) and when both are combined (Full). $\mathcal{S}_{\text{conflict}}$ alone is lightweight, faster than hidden-state baselines such as LUMINA (1.02 / 1.30) and ReDeEP (1.33 / 1.62), while leading them on long-form AUC (0.7999 vs LUMINA 0.7286, ReDeEP 0.7616). $\mathcal{S}_{\text{likelihood}}$ incurs additional decoding cost from candidate and reasoning generation, but drives the short-form AUC gains. Combining both as Full inherits this overhead, which the deployment skip below minimizes for long responses, where $\mathcal{S}_{\text{likelihood}}$'s contribution is already negligible.

Cost controllability. The reasoning budget and candidate set size K control the inference cost of $\mathcal{S}_{\text{likelihood}}$. We use a fixed 20-token reasoning budget and $K = 3$ across all datasets and backbones. Short-form AUC varies mildly across reasoning budgets (Figure 5 right) and remains stable across $K \in \{2, 3, 5\}$ (Table 11); smaller budgets can reduce latency in cost-sensitive settings.

Practical deployment skip. Per Remark in Section 3.3, a practical deployment can omit the bidirectional check in $\mathcal{S}_{\text{likelihood}}$ once $w(T)$ falls below a small threshold where $\mathcal{S}_{\text{likelihood}}$'s contribution to $\mathcal{S}_{\text{final}}$ is already negligible by design. Specifically, we can set the threshold to 0.01, so the bidirectional check is omitted whenever the response is long enough that $w(T) \leq 0.01$. The Deployed row in Table 15 reports the resulting cost: ARGUS runs faster than hidden-state baselines in long-form (0.82 vs LUMINA 1.30, ReDeEP 1.62), where the bidirectional check is skipped. For short-form responses, the bidirectional check remains active because it is the component that drives the short-form accuracy gains; its cost is therefore concentrated in the regime where it is most useful. In our evaluation, this skip preserves AUC because the omitted bidirectional term has negligible weight for long responses: Deployed matches Full's short-form AUC (0.8184) and $\mathcal{S}_{\text{conflict}}$'s long-form AUC (0.7999). In both regimes, these AUCs are higher than those of the baselines in the table.

Method	Short		Long	
	AUC	sec/sample	AUC	sec/sample
SelfCheckGPT	0.6999	1.34	0.6186	11.67
EigenScore	0.7107	1.28	0.6533	8.57
LLMCheck	0.5189	3.47	0.7269	3.60
ReDeEP	0.5827	1.33	0.7616	1.62
LUMINA	0.5782	1.02	0.7286	1.30
ARGUS ($\mathcal{S}_{\text{conflict}}$ only)	0.5229	0.87	0.7999	0.81
ARGUS ($\mathcal{S}_{\text{likelihood}}$ only)	0.8015	2.01	0.3575	8.91
ARGUS (full)	0.8184	2.13	0.7999	9.08
ARGUS (deployed)	0.8184	2.09	0.7999	0.82

Table 15: Per-sample inference time and detection AUC. The first three ARGUS rows decompose the cost of each component. **ARGUS (deployed)** skips the bidirectional check for long responses when $w(T) \leq 0.01$, reducing long-form inference time from 9.08 to 0.82 sec/sample while preserving the reported AUC. In short-form settings, the bidirectional check remains active and accounts for the additional decoding cost.

I Prompt for Reasoning

To generate reasoning for the answers, we used the same prompt uniformly across all short-form and long-form datasets, as shown in Figure 9.

{Query}

{Retrieved Context}

{Model Response}

Refer to the given information and briefly state the reason behind your previous answer/response.

Reason:

Figure 9: Prompt for generating the reasoning.

J Additional Analysis on NQ-swap

While our main evaluation focuses on retrieval-grounded hallucination where the context is accurate, NQ-swap [17] represents a related but distinct setting in which entities in the original NQ contexts are replaced with different entities (e.g., person or location names), so that the context-grounded answer can diverge from the model's parametric knowledge. The faithful response uses the entity from the perturbed context, while responses that follow the original parametric-knowledge entity rather than the perturbed context are treated as hallucination. We include this analysis as a complementary perspective on ARGUS in a perturbed short-form setting.

Performance. Table 16 reports per-method AUC, PRC, and F_1 across the three backbones. ARGUS ranks second in AUC across all three backbones, while achieving the best PRC on two backbones and the second-best PRC on the third.

Why NQ-swap still favors likelihood-based signals. Perplexity directly measures token-level confidence and is well matched to settings where errors manifest as incorrect short final-answer decisions under context–knowledge conflict. Although our bidirectional likelihood is also designed for short-form answer selection, NQ-swap is a more specific case in which the decision is often between the perturbed-context answer and a competing parametric-knowledge answer. Because the post-response reasoning is itself generated by the model, it may still be influenced by parametric knowledge under entity-swapped contexts. As a result, the additional reasoning step provides less extra discrimination beyond forward confidence in NQ-swap than in other short-form settings.

Method	LLaMA2-7B			LLaMA2-13B			LLaMA3-8B		
	AUC	PRC	F_1	AUC	PRC	F_1	AUC	PRC	F_1
Perplexity	**0.7403**	0.8398	0.7878	**0.7646**	0.7948	0.7376	**0.7982**	**0.9376**	0.8964
LN-Entropy	0.6969	0.8111	0.7928	0.7201	0.7509	0.6976	0.7808	0.9287	0.8963
Focus	0.5618	0.7642	0.7944	0.6487	0.7239	0.6992	0.5190	0.8429	0.8991
SelfCKGPT	0.5473	0.6920	0.7943	0.6127	0.6400	0.7165	0.5761	0.8553	0.8961
EigenScore	0.6330	0.7519	0.7808	0.6507	0.6788	0.6793	0.6585	0.8896	0.8793
P(True)	0.6340	0.7843	0.7942	0.3381	0.4636	0.6990	0.6692	0.8938	**0.8992**
RefChecker	0.6345	0.7403	0.7944	0.6289	0.6347	0.6992	0.5912	0.8482	0.8991
ReDeEP	0.6752	0.7933	0.7947	0.5546	0.6004	0.6975	0.7132	0.9240	**0.8992**
LUMINA	0.6659	0.7652	**0.8213**	0.6856	0.6761	**0.7669**	0.5781	0.8751	0.8977
ARGUS (ours)	0.7293	**0.8485**	0.7636	0.7524	**0.8099**	0.7187	0.7877	0.9364	0.8945

Table 16: Performance on NQ-swap, a knowledge-conflict setting with entity-swapped contexts. Best results are in **bold**, and second-best are underlined.

However, Perplexity remains weak in long-form regimes (Table 1), where local fluency can mask gradual drift from the evidence.

Why drift-oriented conflict signals are less effective on NQ-swap. Our hidden-state conflict score $\mathcal{S}_{\text{conflict}}$ is designed to detect drift that accumulates over many tokens, where reliance on internal knowledge progressively grows relative to context grounding. NQ-swap, in contrast, is a short-form setting in which the main error often appears as selecting a parametric-knowledge answer over the perturbed-context answer, rather than as sustained drift over a long generation. In such cases, the model may still remain broadly aligned with the input context at the representation level even when the final answer is wrong, making drift-oriented conflict signals less discriminative. Because NQ-swap responses are also short, the adaptive weighting places greater emphasis on $\mathcal{S}_{\text{likelihood}}$ than on $\mathcal{S}_{\text{conflict}}$ in this regime; this is appropriate, since drift-oriented signals like $\mathcal{S}_{\text{conflict}}$ are less well matched to short-form answer selection under context–knowledge conflict. Notably, this pattern is not unique to ARGUS: in our experiments, other conflict-based detectors such as ReDeEP and LUMINA also underperform Perplexity on NQ-swap, suggesting that drift-oriented conflict signals are less well matched to short-form answer selection in this setting.

How NQ-swap supports our regime-based view. These results reinforce our regime-based view: even when context–knowledge conflict is present, the most informative signal depends on how the failure manifests. In short-form settings, conflict-driven errors often appear as incorrect final-answer decisions between the perturbed-context answer and a competing parametric-knowledge answer, where likelihood-based signals are more discriminative; in long-form settings, conflict more often manifests as sustained drift, which is better captured by hidden-state signals. Response length therefore serves as a practical proxy for the failure regime rather than a direct cause.

K Additional Results on Mistral-7B

To verify that our proposed method generalizes beyond the LLaMA family, we evaluate ARGUS on Mistral-7B [12] against representative output-based (Perplexity) and conflict-based (ReDeEP, LUMINA) baselines from the main results. Table 17 reports performance averaged over the same short-form (6 QA datasets) and long-form (RAGTruth) benchmarks used in Table 1, excluding Dolly as the dataset does not include Mistral-7B responses. As shown in Table 21, RAGTruth alone sufficiently reflects the long-form evaluation, as it exhibits similar relative performance trends across methods.

ARGUS achieves the strongest performance across all six metrics, leading Perplexity in short-form and ReDeEP in long-form. This mirrors the pattern observed on the LLaMA backbones: output-based signals are stronger in short-form, conflict-based signals are stronger in long-form, and adaptive weighting captures both. The result indicates that the regime-aware design is not specific to a particular model family.

Method	Short-form (6 QA datasets)			Long-form (RAGTruth)		
	AUC	PRC	F_1	AUC	PRC	F_1
Perplexity [23]	<u>0.8046</u>	<u>0.8110</u>	<u>0.7562</u>	0.6946	0.6614	0.7233
ReDeEP [26]	0.7816	0.7986	0.7264	<u>0.7938</u>	<u>0.8223</u>	0.6748
LUMINA [31]	0.5567	0.6211	0.6754	0.7261	0.7197	<u>0.7516</u>
ARGUS (ours)	**0.8238**	**0.8118**	**0.7753**	**0.8484**	**0.8606**	**0.7750**

Table 17: Performance on Mistral-7B. Dolly is omitted because its human hallucination annotations exist only for the LLaMA responses released with the benchmark, so no labeled Mistral-7B responses are available.

Algorithm 1 ARGUS Hallucination Detection

Require: Query $\mathbf{q}$, retrieved context $\mathbf{c}$, response $\mathbf{x} = (x_1, \ldots, x_T)$; transition length T_0, top-K, reasoning budget B
Ensure: Hallucination score $\mathcal{S}_{\text{final}}$

// Short-form score (Section 3.1)
$s \leftarrow \left(\prod_{i=1}^{T} p(x_i \mid \mathbf{q}, \mathbf{c}, x_{<i})\right)^{1/T}$ ▷ Forward likelihood, Eq. 1
Generate reasoning $\mathbf{z}$ conditioned on $(\mathbf{q}, \mathbf{c}, \mathbf{x})$ (max B tokens)
Take top-K first-token candidates for x_1 and complete each into $\{\mathbf{x}^{(k)}\}_{k=1}^{K}$ (with $\mathbf{x}^{(1)} = \mathbf{x}$)
$\tilde{s} \leftarrow \frac{s(\mathbf{x}^{(1)}|\mathbf{q},\mathbf{c}) \cdot s(\mathbf{z}|\mathbf{q},\mathbf{c},\mathbf{x}^{(1)})}{\sum_{k=1}^{K} s(\mathbf{x}^{(k)}|\mathbf{q},\mathbf{c}) \cdot s(\mathbf{z}|\mathbf{q},\mathbf{c},\mathbf{x}^{(k)})}$ ▷ Bidirectional likelihood, Eq. 2
$\mathcal{S}_{\text{likelihood}} \leftarrow s \cdot \tilde{s}$ ▷ Combined score, Eq. 3

// Long-form score (Section 3.2)
for $t = 1, \ldots, T$ **do**
  for $\mathrm{s} \in \{\text{context}, \text{query}, \text{response}, \text{ffn}\}$ **do**
    $h_{\mathrm{s}}^{(t)} \leftarrow \text{Norm}\left(\sum_{l=1}^{L} h_{\mathrm{s}}^{(l,t)}\right)$ ▷ Per-source aggregation
  end for
  $S_{\text{ground}}^{(t)} \leftarrow \text{Sim}(h_{\text{context}}^{(t)}, h^{(t)}) + \text{Sim}(h_{\text{query}}^{(t)}, h^{(t)})$ ▷ Eq. 4
  $S_{\text{internal}}^{(t)} \leftarrow \text{Sim}(h_{\text{ffn}}^{(t)}, h^{(t)}) + \text{Sim}(h_{\text{response}}^{(t)}, h^{(t)})$ ▷ Eq. 5
  $\mathcal{S}_{\text{conflict}}^{(t)} \leftarrow S_{\text{internal}}^{(t)} - S_{\text{ground}}^{(t)}$
end for
$\mathcal{S}_{\text{conflict}} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \mathcal{S}_{\text{conflict}}^{(t)}$ ▷ Eq. 6

// Adaptive combination (Section 3.3)
$w(T) \leftarrow 1/(1 + \exp(T - T_0))$ ▷ Length-based weight
$\mathcal{S}_{\text{final}} \leftarrow (1 - w(T)) \cdot \mathcal{S}_{\text{conflict}} - w(T) \cdot \mathcal{S}_{\text{likelihood}}$ ▷ Final score, Eq. 7
return $\mathcal{S}_{\text{final}}$

L ARGUS Pseudo-code

Algorithm 1 provides the pseudo-code of ARGUS.

M Bootstrap Confidence Intervals

We quantify evaluation-set uncertainty conditional on the fixed responses and detection scores using paired bootstrap resampling on LLaMA2-13B. For each of 1,000 resamples, we draw samples with replacement within each dataset, using the same indices for both methods, compute the per-dataset AUC of each method, average over datasets to obtain macro AUC, and record the difference; the 95%

confidence interval is the 2.5th to 97.5th percentile of these differences. Table 18 reports the results for ARGUS against selected baselines in each regime.

Setting	Comparison	ΔAUC	95% CI
Short-form (6 QA datasets)	ARGUS vs Perplexity	+0.056	[+0.052, +0.059]
Long-form (RAGTruth, Dolly)	ARGUS vs LUMINA	+0.071	[+0.003, +0.136]
Long-form (RAGTruth, Dolly)	ARGUS vs ReDeEP	+0.038	[−0.021, +0.103]
Long-form (RAGTruth only)	ARGUS vs ReDeEP	+0.033	[+0.004, +0.062]

Table 18: Paired bootstrap confidence intervals for the main comparisons.

The short-form gain over Perplexity and the long-form gain over LUMINA have confidence intervals above zero. The long-form gain over ReDeEP on the combined set has an interval that includes zero, while on RAGTruth alone it excludes zero; we therefore do not interpret the combined-set difference from ReDeEP as conclusive.

N Full Per-Dataset Results

This appendix reports per-dataset breakdowns of the aggregated results in Table 1 of the main paper. Tables 19 and 20 cover the six short-form QA datasets (HotpotQA, TriviaQA, SearchQA, SQuAD, NewsQA, NQ); Table 21 covers the two long-form benchmarks (RAGTruth, Dolly).

O Limitations and Future Work

ARGUS is a white-box method, requiring access to both token-level probabilities and hidden representations. Like other detectors that access model internals (ReDeEP, LUMINA, EigenScore), it is applicable to open-source LLM deployments; pure black-box scenarios (text-only API) are typically addressed by methods that require only the generated text, such as SelfCheckGPT and RefChecker. Additionally, ARGUS focuses on hallucination detection; using ARGUS scores to trigger mitigation actions (e.g., regeneration or refusal) is a complementary direction left for future work.

P Broader Impact

ARGUS targets a practical reliability problem in retrieval-augmented generation: distinguishing hallucinated outputs from faithful ones. Reliable detection is a prerequisite for deploying RAG systems in high-stakes settings (e.g., medical or legal question-answering), where unfounded answers can mislead users. Our detector is lightweight, especially in the deployed variant, making it practical to integrate into existing RAG pipelines.

Our regime-aware design uses response length as a practical proxy for the typical failure pattern in each regime (candidate misselection vs. gradual drift), which captures the dominant patterns in real-world RAG use. We encourage integrating ARGUS scores within broader system-level safety practices.

Model	Method	HotpotQA (Short)			TriviaQA (Short)			SearchQA (Short)		
		AUC	PRC	F_1	AUC	PRC	F_1	AUC	PRC	F_1
LLaMA2-7B	Perplexity	0.7831	0.7772	0.7708	0.7388	0.6913	0.6778	0.7878	0.7231	0.7251
	LN-Entropy	0.7502	0.7509	0.7446	0.7145	0.6669	0.6624	0.7467	0.6812	0.6989
	Focus	0.6153	0.6974	0.7101	0.6093	0.6000	0.6181	0.5664	0.5625	0.6282
	SelfCKGPT	0.6630	0.6899	0.6733	0.7099	0.6494	0.6659	0.7229	0.6525	0.6667
	EigenScore	0.6858	0.6830	0.7228	0.6651	0.6253	0.6426	0.6616	0.5657	0.6710
	P(True)	0.5698	0.6608	0.7072	0.6543	0.6209	0.6118	0.5578	0.5287	0.6282
	RefChecker	0.5398	0.5783	0.7120	0.5167	0.4673	0.6181	0.5090	0.4695	0.6282
	LLMCheck	0.4958	0.5524	0.7111	0.4682	0.4324	0.6178	0.4854	0.4546	0.6281
	ReDeEP	0.5874	0.6438	0.7118	0.6382	0.6205	0.6183	0.6817	0.6610	0.6285
	LUMINA	0.5346	0.5970	0.7115	0.5030	0.4858	0.6179	0.5835	0.5463	0.6331
	ARGUS (ours)	**0.8181**	**0.8250**	**0.7858**	**0.7628**	**0.7222**	**0.6927**	**0.8158**	**0.7529**	**0.7388**
LLaMA2-13B	Perplexity	0.8077	0.7485	0.7385	0.7264	0.6513	**0.6520**	0.7733	0.6374	0.6451
	LN-Entropy	0.7779	0.7210	0.7229	0.6954	0.6181	0.6342	0.7283	0.5838	0.6173
	Focus	0.6506	0.6638	0.6004	0.6133	0.5685	0.5894	0.5805	0.4726	0.5378
	SelfCKGPT	0.6897	0.6668	0.6465	0.7070	0.6378	0.6355	0.7261	0.5890	0.6187
	EigenScore	0.6997	0.6324	0.6927	0.6878	0.6055	0.6195	0.6744	0.4835	0.6163
	P(True)	0.4148	0.4275	0.6434	0.4805	0.4069	0.5835	0.4567	0.3420	0.5375
	RefChecker	0.5300	0.4952	0.6455	0.5179	0.4362	0.5894	0.5111	0.3804	0.5379
	LLMCheck	0.5153	0.4856	0.6423	0.4802	0.4135	0.5886	0.4901	0.3673	0.5377
	ReDeEP	0.5285	0.5022	0.6432	0.5370	0.4422	0.5891	0.5284	0.3916	0.5421
	LUMINA	0.5080	0.5047	0.6444	0.5175	0.4676	0.5892	0.6036	0.4977	0.5384
	ARGUS (ours)	**0.8279**	**0.8041**	**0.7467**	**0.7430**	**0.6820**	0.6489	**0.8044**	**0.6822**	**0.6676**
LLaMA3-8B	Perplexity	0.6911	0.7292	0.7687	0.7120	0.6656	0.6881	0.7362	0.6241	0.6246
	LN-Entropy	0.6786	0.7284	0.7547	0.6904	0.6565	0.6718	0.7039	0.5925	0.6087
	Focus	0.5697	0.6949	0.7500	0.6285	0.6177	0.6431	0.6333	0.5382	0.5291
	SelfCKGPT	0.6474	0.7263	0.7571	0.6859	0.6605	0.6706	0.7328	0.6235	0.6196
	EigenScore	0.6577	0.7064	0.7372	0.7100	0.6476	0.6844	0.6878	0.5066	0.6285
	P(True)	0.7090	**0.7973**	0.7265	0.6490	0.6347	0.6424	0.6562	0.5628	0.5502
	RefChecker	0.5723	0.6589	0.7528	0.5219	0.4971	0.6438	0.5107	0.3937	0.5536
	LLMCheck	0.4904	0.6014	0.7471	0.4824	0.4680	0.6427	0.4950	0.3913	0.5534
	ReDeEP	0.6777	0.7776	0.6894	0.5969	0.5635	0.6385	0.5917	0.5222	0.5505
	LUMINA	0.5938	0.7024	0.7446	0.5146	0.5078	0.6440	0.5397	0.4307	0.5520
	ARGUS (ours)	**0.7263**	0.7732	**0.7769**	**0.7383**	**0.7035**	**0.6917**	**0.7497**	**0.6390**	**0.6365**

Table 19: Hallucination detection performance on HotpotQA, TriviaQA, and SearchQA. Best results are in **bold**, and second-best are underlined.

Model	Method	SQuAD (Short)			NewsQA (Short)			NQ (Short)		
		AUC	PRC	F_1	AUC	PRC	F_1	AUC	PRC	F_1
LLaMA2-7B	Perplexity	0.6956	0.7487	0.7631	0.7305	0.8129	0.8185	0.6996	0.7780	0.7700
	LN-Entropy	0.6672	0.7308	0.7537	0.7096	0.8120	0.8095	0.6294	0.7336	0.7671
	Focus	0.5108	0.6744	0.7507	0.5185	0.7379	0.8026	0.5568	0.7314	0.7678
	SelfCKGPT	0.6394	0.6965	0.7652	0.6980	0.8203	0.8125	0.6224	0.7110	0.7800
	EigenScore	0.6978	0.7325	0.7571	0.6918	0.7599	0.8157	0.7578	0.7937	0.8053
	P(True)	0.7731	**0.8421**	0.7622	0.6990	0.8289	0.7772	0.7127	0.8123	0.7666
	RefChecker	0.5549	0.6428	0.7506	0.5187	0.6798	0.8026	0.5490	0.6546	0.7678
	LLMCheck	0.4987	0.5955	0.7462	0.5316	0.6914	0.8027	0.5702	0.6883	0.7718
	ReDeEP	0.6883	0.7503	0.3428	0.6830	0.7970	0.6414	0.6990	0.8018	0.7681
	LUMINA	0.5767	0.6496	0.7538	0.6247	0.7637	0.7944	0.5951	0.6807	0.7738
	ARGUS (ours)	**0.7888**	0.8096	**0.7999**	**0.8109**	**0.8765**	**0.8461**	**0.8500**	**0.8974**	**0.8466**
LLaMA2-13B	Perplexity	0.7648	0.7134	0.7133	0.7984	0.8721	0.8403	0.7072	0.7566	0.7521
	LN-Entropy	0.7261	0.6825	0.6969	0.8068	0.8938	0.8450	0.6442	0.7151	0.7413
	Focus	0.5698	0.6133	0.6513	0.4780	0.7362	0.8179	0.6098	0.7341	0.7425
	SelfCKGPT	0.6809	0.6509	0.6805	0.7881	0.9058	0.8281	0.6075	0.6749	0.7529
	EigenScore	0.7583	0.6916	0.7172	0.7048	0.7946	0.8320	0.7392	0.7641	0.7701
	P(True)	0.3396	0.3911	0.6504	0.3425	0.6076	0.8177	0.2748	0.4649	0.7424
	RefChecker	0.5325	0.5117	0.6516	0.5038	0.6940	0.8180	0.5493	0.6232	0.7419
	LLMCheck	0.5145	0.4879	0.6512	0.5193	0.6996	0.8179	0.5938	0.6861	0.7449
	ReDeEP	0.5754	0.5371	0.6455	0.6785	0.8531	0.8171	0.6481	0.7266	0.7376
	LUMINA	0.5656	0.5174	0.6487	0.6719	0.8059	0.8166	0.6025	0.6545	0.7475
	ARGUS (ours)	**0.8258**	**0.7917**	**0.7637**	**0.8701**	**0.9257**	**0.8758**	**0.8392**	**0.8824**	**0.8343**
LLaMA3-8B	Perplexity	0.5849	0.7629	0.8186	0.6166	0.7805	0.8329	0.6793	0.8552	0.8572
	LN-Entropy	0.5812	0.7607	0.8209	0.6008	0.7748	0.8319	0.6243	0.8342	0.8575
	Focus	0.4968	0.7358	0.8287	0.5421	0.7718	0.8324	0.5814	0.8203	0.8583
	SelfCKGPT	0.5949	0.7765	0.8283	0.6923	**0.8462**	0.8331	0.6095	0.8199	0.8546
	EigenScore	0.6269	0.7775	0.8273	0.6933	0.7963	**0.8431**	0.7378	0.8712	0.8694
	P(True)	**0.7269**	**0.8412**	**0.8351**	0.5025	0.7425	0.8297	0.7313	0.8863	0.8584
	RefChecker	0.5565	0.7412	0.8284	0.5158	0.7212	0.8321	0.5434	0.7736	0.8589
	LLMCheck	0.4714	0.6931	0.8282	0.5293	0.7248	0.8323	0.5554	0.7808	0.8527
	ReDeEP	0.6439	0.8071	0.8284	0.6325	0.8143	0.8329	0.7002	0.8875	0.8587
	LUMINA	0.5294	0.7348	0.8309	0.6002	0.7877	0.8275	0.5533	0.8070	0.8576
	ARGUS (ours)	0.6494	0.8075	0.8330	**0.6957**	0.8400	0.8346	**0.8029**	**0.9079**	**0.8882**

Table 20: Hallucination detection performance on SQuAD, NewsQA, and NQ. Best results are in **bold**, and second-best are underlined.

Model	Method	RAGTruth (Long)			Dolly (Long)		
		AUC	PRC	F_1	AUC	PRC	F_1
LLaMA2-7B	Perplexity	0.5146	0.4903	0.6717	0.3133	0.4780	0.7389
	LN-Entropy	0.6965	0.6631	0.6768	0.6699	0.7317	0.7500
	Focus	0.5911	0.5852	0.6627	0.6106	0.7190	0.7465
	SelfCKGPT	0.5034	0.5217	0.6696	0.7107	0.7530	0.7642
	EigenScore	0.5271	0.5092	0.6686	0.6455	0.7366	<u>0.7703</u>
	P(True)	0.5535	0.5310	0.6637	**0.7174**	**0.7834**	0.7500
	RefChecker	0.5838	0.5479	0.6686	0.5690	0.6678	0.7389
	LLMCheck	0.6718	0.6140	0.6696	<u>0.7136</u>	<u>0.7549</u>	**0.7794**
	ReDeEP	0.7325	0.7010	0.6696	0.6791	0.7426	0.7586
	LUMINA	**0.7616**	<u>0.7353</u>	<u>0.6894</u>	0.5294	0.6447	0.7389
	ARGUS (ours)	<u>0.7615</u>	**0.7568**	**0.7084**	0.6173	0.7269	0.7550
LLaMA2-13B	Perplexity	0.4649	0.4082	0.6426	0.2996	0.4461	0.7059
	LN-Entropy	0.7675	0.6874	0.6974	0.6202	0.6750	0.7222
	Focus	0.6958	0.6243	0.6348	0.6885	0.7317	0.7463
	SelfCKGPT	0.4585	0.4200	0.6372	**0.7786**	**0.8424**	0.7679
	EigenScore	0.5963	0.5494	0.6349	0.7103	0.6890	<u>0.7680</u>
	P(True)	0.6567	0.6415	0.6248	0.7251	<u>0.7625</u>	0.7563
	RefChecker	0.6546	0.5302	0.6301	0.5877	0.6321	0.7059
	LLMCheck	0.7381	0.6254	0.7026	0.7156	0.7545	0.7376
	ReDeEP	0.8157	0.7987	<u>0.7402</u>	0.7074	0.7386	0.7481
	LUMINA	<u>0.8456</u>	<u>0.8039</u>	0.7185	0.6115	0.6624	0.7571
	ARGUS (ours)	**0.8487**	**0.8350**	**0.7619**	<u>0.7510</u>	0.7433	**0.7937**
LLaMA3-8B	Perplexity	0.7132	0.7203	<u>0.7289</u>	0.6156	0.6152	0.6030
	LN-Entropy	0.7072	0.7133	0.7233	0.6459	0.6482	0.5941
	Focus	0.6231	0.6376	0.7120	0.6156	0.5264	0.6136
	SelfCKGPT	0.5198	0.5499	0.7217	0.6690	0.4170	0.6080
	EigenScore	0.5618	0.6182	0.7097	0.6396	0.5374	0.6286
	P(True)	0.4968	0.5505	0.7048	**0.8070**	**0.7784**	**0.7158**
	RefChecker	0.5882	0.5881	0.7013	0.5597	0.4798	0.5857
	LLMCheck	0.6313	0.6271	0.7124	<u>0.7283</u>	0.6874	0.6667
	ReDeEP	**0.7522**	0.7836	0.7222	0.6413	0.5789	0.6429
	LUMINA	<u>0.7514</u>	**0.7937**	**0.7295**	0.6955	0.6659	0.6379
	ARGUS (ours)	0.7367	<u>0.7900</u>	0.7100	0.7191	<u>0.7119</u>	<u>0.6882</u>

Table 21: Hallucination detection performance on RAGTruth and Dolly. Best results are in **bold**, and second-best are <u>underlined</u>.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims and contributions are clearly presented in the abstract and introduction, and are supported by comprehensive experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of our work in Appendix O.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental settings are described in Section 4, with additional implementation details, datasets, and the algorithm in Appendices A, B, and L, respectively.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets, with all relevant details described in Section 4.1. The code is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details, including hyperparameters, datasets, and threshold selection, are provided in Section 4.1 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report 95% confidence intervals for the AUC differences between ARGUS and selected baselines in each regime on LLaMA2-13B, obtained by paired bootstrap resampling over the evaluation set (Appendix M).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware setup and runtime measurement protocol are described in Appendix A, with detailed complexity and inference-time analysis in Appendix H. The effect of reasoning token budget on detection performance is also studied in Section 4.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and our research conforms to its principles.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix P.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release any new high-risk models or datasets; our work uses publicly available models and datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, models, and baselines are properly cited in Section 4.1 and Appendices B and A.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Code is released as supplementary material with implementation details in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our method is applied to pre-trained LLMs as backbone models. In the proposed bidirectional likelihood score, the same backbone LLM generates a short reasoning sequence conditioned on the query, retrieved context, and response; this sequence is then used for detection scoring. The generation procedure and implementation details are described in Section 3.1 and Appendices A and I.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.